Supplemental Material

Contents

S1 Annotation workflow details **2**

S2 ICU-REACT dataset details **2**
S2.1 ICU-REACT-Train . . . 2
S2.2 ICU-REACT-Test . . . 2
S2.3 Evaluation rubrics . . . 5

S3 Training hyperparameters **6**

S4 Comparisons against all baselines **7**

S5 Ablations **11**
S5.1 Training tasks . . . 12
S5.2 Task ablation . . . 12
S5.3 Dataset size ablation . . . 15

S6 Comparison with Frontier Models **16**

S7 Performance stratified by clinical content **19**

S8 Information seeking performance by data category **22**

S9 Multiple choice medical benchmarks vs clinical reasoning **25**

S10 Data contamination analysis **28**

S11 ICU Variable and Taxonomy Framework **28**

S12 ICU Topics Dimensions **38**

S13 Prompts **48**
S13.1 Seed generation prompts . . . 48
S13.2 Dataset augmentation prompts . . . 65
S13.3 Training prompts . . . 69
S13.4 Inference prompts . . . 71
S13.5 Evaluation prompts . . . 84

References **92**

S1 Annotation workflow details

To ensure the clinical validity of ICU-REACT, candidate items were reviewed by a team of clinicians through a web-based annotation tool (Fig. S1). After authenticating, an annotator was shown the clinical context and question for a single item, along with its clinical-domain tags, and could either begin the annotation or skip items outside their expertise. The review then proceeded through five sequential stages. In Question Validity, the annotator judged whether the question was one they would plausibly ask when caring for a patient in the given context. If the annotator considered the question to be not relevant, then annotation for the item was terminated and would proceed to the next item. In Required Data Elements, they selected the variables needed to answer the question from category-organized toggles (e.g., cardiac imaging results, renal function labs, fluid balance records, and vital signs). In Missing Data Elements, they assessed whether the item's retrieved data captured everything they would query in an EHR and, when necessary, added further variables by searching the OMOP-based variable taxonomy by name or concept ID, or by entering custom variables. In Reasoning, they judged whether the accompanying rationale linking the selected data to the clinical decision was sound. Finally, in Critical Feedback, they could leave free-text suggestions to improve the item even when all prior checks passed. Such written feedback could be as detailed or concise as the annotator wished. Completed annotations were persisted to a MongoDB database and returned to the research team, who aggregated the clinician judgments to filter, correct, and finalize items for the ICU-REACT seed dataset.

S2 ICU-REACT dataset details

S2.1 ICU-REACT-Train

ICU-REACT-Train-Small (Clin-REACT 8B) splits its 10,000 examples evenly into 5,000 reasoning-refinement and 5,000 variable-selection examples, pairing the reasoning task with explicit information-retrieval supervision at the smallest scale, whereas ICU-REACT-Train-Medium (5,307 examples; Clin-REACT 14B and 31B) and ICU-REACT-Train-Large (27,973 examples; Clin-REACT 70B) consist solely of reasoning-refinement examples. Each such example is a two-stage target, in which an initial reasoning trace is revised into an improved trace that the model learns to produce. Figure S2 shows t-SNE projections of the three sets colored by primary ICU topic: for each scale, the left panel depicts the full set of input–output examples, while the middle and right panels separate the initial- and improved-reasoning stages, illustrating the overall sample distribution and its shift between stages.

S2.2 ICU-REACT-Test

The full topic distribution of ICU-REACT-Test (Fig. S3B) is respiratory failure ($n = 16$), hemodynamic instability/shock ($n = 11$), renal failure/electrolyte disorders ($n = 9$), nutrition and metabolic support ($n = 8$), hematologic/coagulation issues ($n = 8$), sedation, pain, and delirium management ($n = 7$), sepsis and severe infections ($n = 6$),

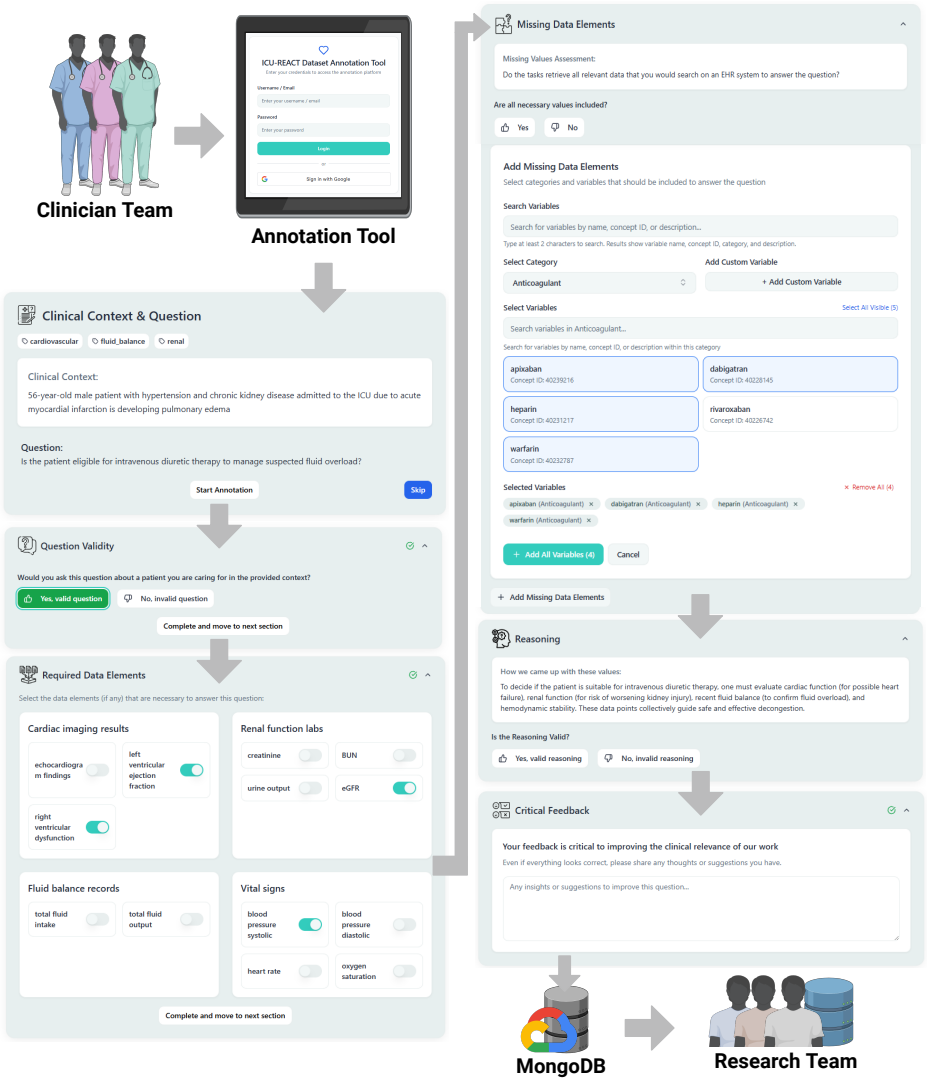


Fig. S1 Clinician annotation workflow and interface for the ICU-REACT seed dataset. The panels show screen captures of the web-based annotation tool for a single example being annotated, with arrows indicating the order of the workflow. A clinician team logs in and is presented with the Clinical Context and Question, tagged by relevant clinical domains, and chooses to annotate or skip the item. The annotation then proceeds through sequential sections: (i) Question Validity, confirming whether the question is one a clinician would ask for such a patient; (ii) Required Data Elements, toggling the variables needed to answer the question, organized by category; (iii) Missing Data Elements, indicating whether the retrieved data are complete and, if not, adding further variables by name or concept ID or as custom entries; (iv) Reasoning, judging whether the provided rationale is valid; and (v) Critical Feedback, an open-text field for additional suggestions. Completed annotations are written to a MongoDB store and passed to the research team for use in the construction of the ICU-REACT seed dataset. Note: the example shown is for illustration of the interface only and is not a real nor validated annotated sample.

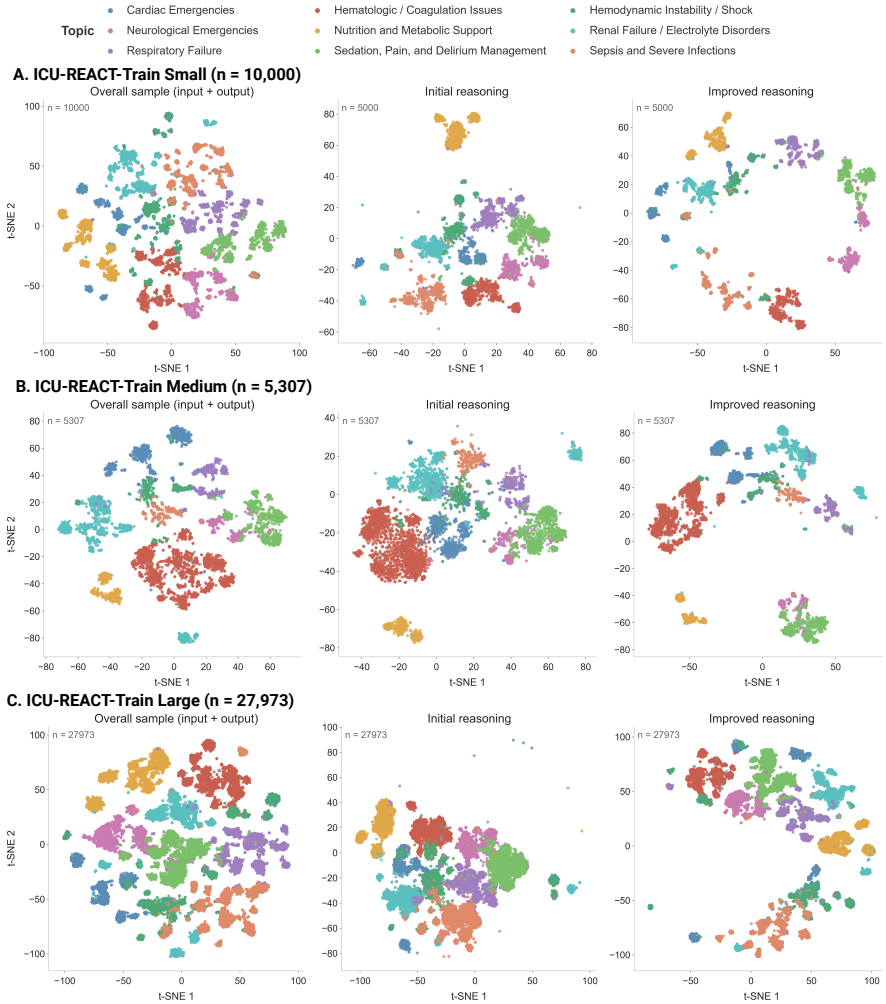


Fig. S2 Topic distribution of the ICU-REACT-Train datasets. t-distributed stochastic neighbor embedding (t-SNE) projections of the ICU-REACT-Train Small (A), Medium (B), and Large (C) datasets, with samples colored according to the primary ICU topic represented in each question. Dataset names (Small, Medium, Large) refer to the parameter scale of the target model each set was used to train. For each dataset scale, the left panel depicts the overall distribution of input-output training examples, whereas the middle and right panels separately visualize examples corresponding to the initial-reasoning and improved-reasoning stages.

neurological emergencies ($n = 4$), and cardiac emergencies ($n = 2$), with a two-dimensional t-SNE projection of the question embeddings showing coherent clustering by topic. Ordered by prevalence, the top-level information categories are physiology (92%), laboratory measurements (75%), medications (66%), scores and assessments (49%), imaging (48%), intervention (46%), diagnosis (44%), and administrative (10%) (Fig. S3C); within each, prevalence concentrated in a few dominant sub-categories,



Fig. S3 ICU-REACT-Test dataset description. (A) Summary of dataset composition, including the number of samples and clinical topics, and the median number of categories, sub-categories, and variables represented per question. (B) t-distributed stochastic neighbor embedding (t-SNE) visualization of questions according to their clinical content, colored by the nine ICU topic areas, alongside the number of samples on each topic. (C) Proportion of test-set questions containing each information category, grouped into physiologic variables, laboratory measurements, medications, scores and assessments, imaging, interventions, diagnoses, and administrative information.

including vitals (82%) within physiology, the comprehensive metabolic panel (61%) within laboratory measurements, and pressors (25%) within medications.

S2.3 Evaluation rubrics

Model outputs on ICU-REACT-Test were scored by an LLM-Judge against item-level rubrics (Fig. S4). The rubric set covers all 71 test samples with 845 items across five evaluation dimensions, at a median of 12.0 items per sample (IQR 11.0–13.0), a median maximum of 144.0 score points per sample (IQR 138.0–153.0), and a median description length of 44.0 words (IQR 40.0–48.0) (Fig. S4A). Items are dominated by reasoning correctness (73.6%, $n = 622$), followed by safety, reasoning synthesis, and task faithfulness (each 8.4%, $n = 71$), and critical anchor (1.2%, $n = 10$) (Fig. S4B).

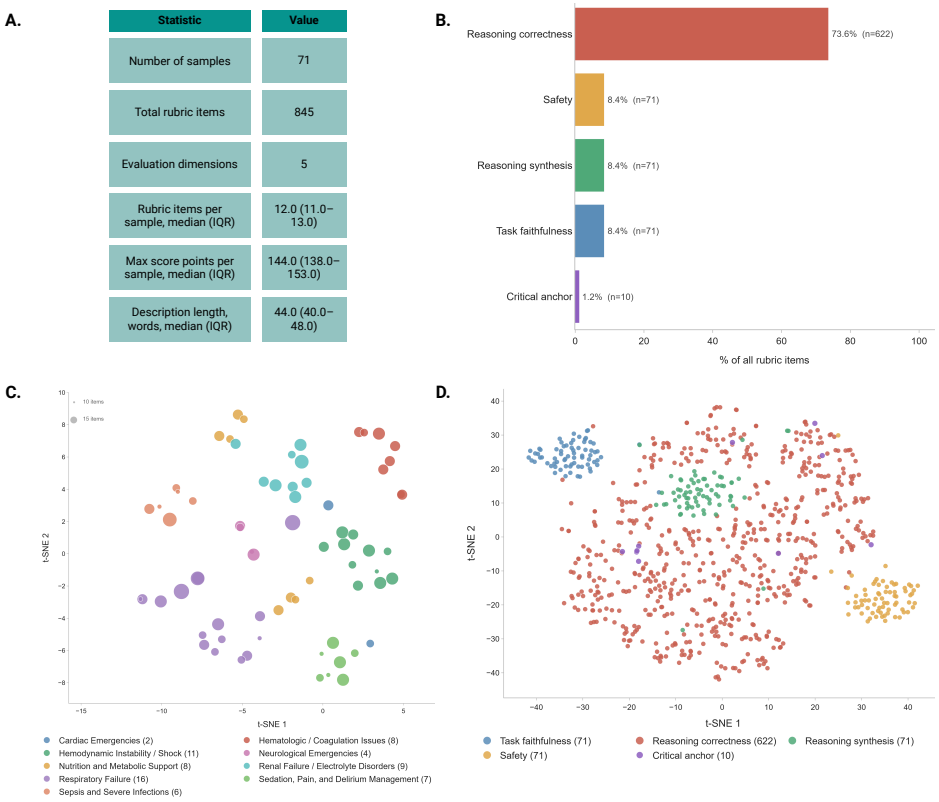


Fig. S4 Rubric statistics and item distribution for the LLM-Judge evaluation on the ICU-REACT test set. (A) Summary statistics of the rubric set, comprising 71 samples and 845 rubric items across 5 evaluation dimensions, together with the per-sample median (IQR) of rubric items, maximum score points, and rubric-description length in words. (B) Distribution of rubric items across the five evaluation dimensions (Reasoning correctness, Safety, Reasoning synthesis, Task faithfulness, and Critical anchor), expressed as a percentage of all 845 rubric items with item counts in parentheses. (C) Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) projection at the sample level, in which each marker is one of the 71 samples, sized by its number of rubric items and colored by its ICU-REACT clinical topic; legend counts give the number of samples per topic. (D) Two-dimensional t-SNE projection at the rubric-item level, in which each point is one of the 845 rubric items, colored by its evaluation dimension; legend counts give the number of items per dimension.

Two-dimensional t-SNE projections color the items by their source sample's clinical topic (Fig. S4C) and by evaluation dimension (Fig. S4D), with both topics and dimensions forming visibly separated clusters.

S3 Training hyperparameters

All Clin-REACT variants were fine-tuned with low-rank adaptation (LoRA), leaving the backbone weights frozen and updating only the injected adapters. We used identical adapter settings across all four scales ($r = 16$, $\alpha = 32$, dropout 0.05); in preliminary experiments, varying the rank, scaling factor, and dropout produced no

appreciable change in downstream performance, so a single configuration was retained across scales to keep adapter capacity constant. Optimization used a learning rate of 1×10^{-5} with a weight decay of 0.01 and a linear warmup over the first 100 steps. All models were trained with a per-device batch size of 4 and 4 gradient accumulation steps, giving an effective batch size of 16. Clin-REACT 8B was trained for 5 epochs and the remaining variants for 3. Each training corpus was split 90:10 into training and validation partitions, with the validation partition used only for checkpoint selection. Table S1 reports the full configuration for each variant.

Table S1 Training hyperparameters for each Clin-REACT variant.

Hyperparameter	Clin-REACT 8B	Clin-REACT 14B	Clin-REACT 31B	Clin-REACT 70B
LoRA rank (r)	16	16	16	16
LoRA alpha (α)	32	32	32	32
LoRA dropout	0.05	0.05	0.05	0.05
Learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Weight decay	0.01	0.01	0.01	0.01
Warmup steps	100	100	100	100
Batch size (per device)	4	4	4	4
Gradient accumulation steps	4	4	4	4
Effective batch size	16	16	16	16
Epochs	5	3	3	3
Training samples (total)	10,000	5,307	5,307	27,973
Training samples (train split)	9,000	4,776	4,776	25,176
Train:validation split (%)	90:10	90:10	90:10	90:10

S4 Comparisons against all baselines

A detailed comparison of each Clin-REACT model against all baseline models can be seen in Fig. S5. When compared with its backbone architecture (Llama 3.1 8B Instruct), Clin-REACT 8B achieved statistically significant gains of +12.5% ($p < 0.001$), +11.0% ($p < 0.01$), +7.5% ($p < 0.001$) and +2.7% ($p < 0.001$) on ICU-REACT, SCT-Bench, ER-Reason, and VivaBench, respectively, with slight degradation on MedRBench (-3.9%, $p < 0.001$). On the other hand, Clin-REACT 14B improved across all benchmarks with respect to its backbone (Baichuan M1 14B Instruct), with mostly significant gains of +7.4% ($p < 0.001$), +3.3% ($p > 0.05$), +7.5% ($p < 0.001$), +3.6% ($p < 0.001$) and +1.8% ($p < 0.05$) on ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench, respectively.

Similarly, Clin-REACT 31B outperformed its backbone (Gemma 4 31B) on four of five benchmarks, with statistically significant gains of +4.2% on ICU-REACT ($p < 0.001$), +3.9% on ER-Reason ($p < 0.001$), +2.9% on MedRBench ($p < 0.001$), and +1.9% on VivaBench ($p < 0.05$). Although performance decreased slightly on SCT-Bench (-2.1%), this difference was not statistically significant ($p > 0.05$). Clin-REACT 70B demonstrated the largest improvements over its backbone (Llama 3.3 70B Instruct), achieving gains of +14.1% ($p < 0.001$), +7.6% ($p < 0.01$), +5.6% ($p < 0.001$), and +1.0% ($p < 0.001$) on ICU-REACT, SCT-Bench, ER-Reason, and

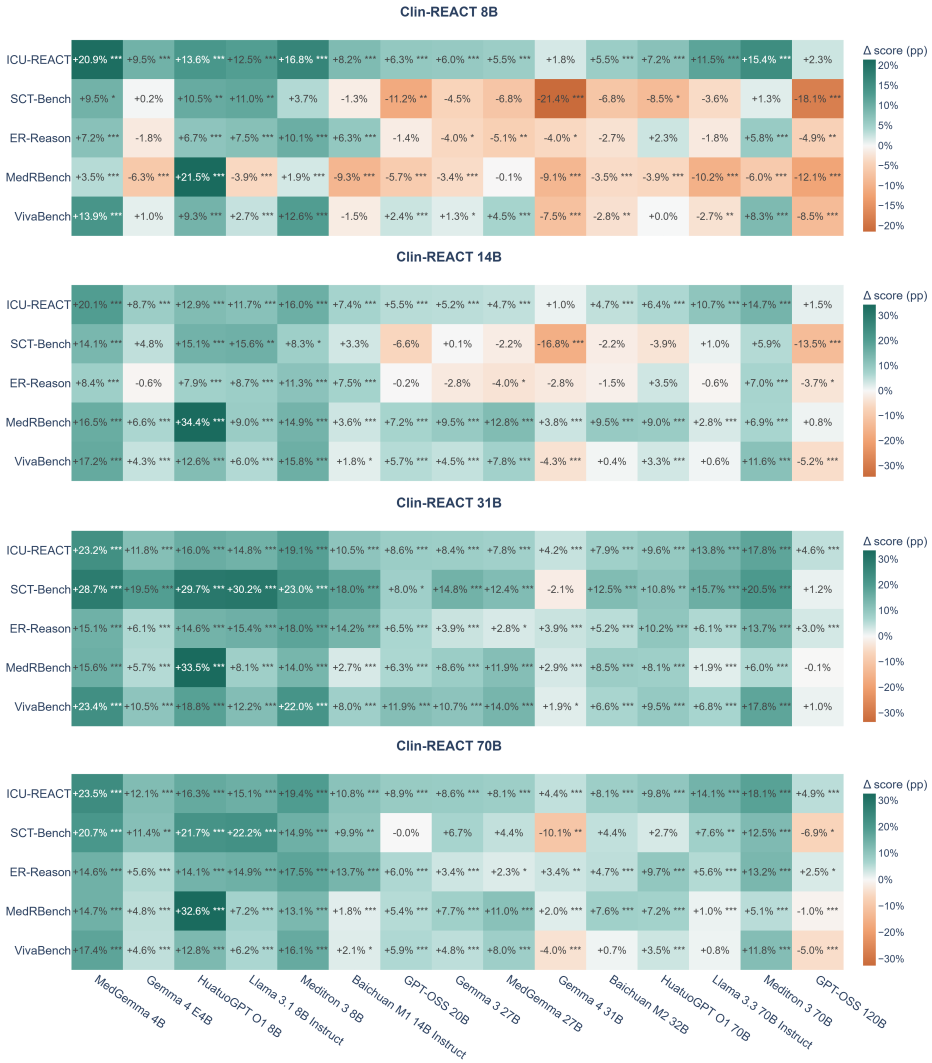


Fig. S5 Per-benchmark score differences between each Clin-REACT model and every baseline model across five clinical-reasoning benchmarks. Panels correspond to the four Clin-REACT models (top to bottom: Clin-REACT 8B, 14B, 31B, and 70B). Each cell reports the difference in score (Δ, percentage points) between the Clin-REACT model and a baseline model on one benchmark (rows), computed as the Clin-REACT model minus the baseline using each benchmark's average-across-metrics score; green indicates the Clin-REACT model scored higher, orange indicates the baseline scored higher, and color intensity represents the magnitude of the difference. Columns are the 15 open-source baseline models (medical and general purpose), ordered by parameter count; rows are the five clinical-reasoning benchmarks (ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench). Numeric labels give each cell's Δ, and asterisks denote the statistical significance of the paired difference (two-sided Wilcoxon signed-rank test; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$). Each panel uses an independent color scale.

Table S2 Performance of Clin-REACT models against baselines across clinical reasoning benchmarks.

Model	ICU-REACT			SCT-Bench	ER-Reason			MedR-Bench				VivaBench		
	Parent F1	Variable F1	Reasoning	SCT	Decision	Differential	Treatment	Assess. Rec.	Assess. Prec.	Diagnosis	Treatment	Overall Recall	Overall Precision	Final Dx
Clin-REACT 70B	42.7 (39.4–46.3)	31.6 (28.9–34.3)	**60.9 (58.2–63.4)*****	67.5 (61.6–73.0)	58.4 (52.9–63.5)	52.2 (46.2–57.9)	**41.9 (36.3–47.3)**	47.7 (46.1–49.6)	25.8 (24.7–27.1)	65.2 (62.1–68.3)	45.4 (41.1–49.8)	30.3 (29.1–31.4)	18.3 (17.6–19.0)	34.0 (31.2–37.0)
Clin-REACT 31B	**45.1 (42.0–48.3)**	**35.4 (32.9–37.9)**	53.8 (51.7–55.8)	75.5 (69.8–80.6)	**60.1 (54.2–65.7)***	**54.2 (48.3–60.7)**	39.7 (34.1–45.3)	48.0 (46.2–49.7)	29.8 (28.4–31.1)	66.0 (62.7–69.0)	44.0 (39.8–48.3)	29.5 (28.4–30.6)	22.8 (21.9–23.7)	**48.2 (45.4–51.5)**
Clin-REACT 14B	38.9 (35.5–42.6)	25.8 (23.2–28.3)	60.2 (57.8–62.8)	60.9 (54.7–66.8)	51.4 (45.5–57.3)	47.2 (42.1–52.7)	35.4 (30.1–40.9)	48.5 (46.9–50.3)	29.5 (28.2–30.6)	59.1 (55.7–62.2)	**54.4 (50.0–58.5)*****	27.3 (26.3–28.4)	21.7 (20.8–22.7)	32.9 (29.7–35.9)
Clin-REACT 8B	41.8 (38.7–44.7)	34.1 (31.5–36.8)	51.4 (49.1–53.6)	56.3 (49.7–62.2)	55.8 (50.9–60.6)	44.8 (39.9–49.7)	29.8 (25.4–34.5)	42.6 (41.0–44.3)	23.5 (22.3–24.8)	48.5 (45.1–52.0)	25.1 (21.6–29.3)	26.6 (25.4–27.6)	20.3 (19.5–21.1)	25.3 (22.3–28.2)
HuatuoGPT O1 70B	42.1 (39.3–45.1)	29.6 (27.2–32.2)	33.8 (32.3–35.1)	64.8 (58.6–71.2)	50.4 (45.0–55.7)	40.4 (34.3–46.5)	32.7 (27.2–38.2)	29.8 (28.2–31.4)	**49.0 (47.0–51.3)*****	47.5 (44.4–50.8)	28.8 (24.9–33.0)	14.1 (13.3–14.9)	27.9 (26.3–29.3)	30.0 (27.1–33.1)
Meditron 3 70B	36.0 (31.7–39.8)	19.9 (16.7–22.9)	25.0 (23.5–26.5)	55.0 (48.8–61.9)	43.0 (37.6–48.2)	43.0 (37.2–49.4)	26.9 (21.5–32.3)	36.6 (34.9–38.3)	37.1 (35.5–38.7)	53.7 (50.5–57.0)	36.3 (32.4–41.1)	14.8 (13.9–15.8)	22.4 (21.0–23.7)	10.1 (8.1–11.9)
Baichuan M2 32B	41.1 (38.6–43.9)	31.3 (29.1–33.4)	38.3 (36.6–40.0)	63.0 (56.4–68.8)	57.4 (52.2–62.8)	49.0 (43.3–54.5)	32.1 (27.2–37.1)	48.0 (46.4–49.7)	16.7 (15.9–17.6)	55.5 (52.0–58.9)	33.4 (29.0–38.2)	29.4 (28.3–30.6)	17.8 (17.1–18.6)	33.4 (30.4–36.6)
MedGemma 27B	41.5 (38.6–44.4)	30.0 (27.5–32.4)	39.3 (37.3–41.3)	63.1 (56.8–68.9)	57.7 (51.6–63.4)	53.0 (46.8–59.7)	35.2 (30.1–40.5)	32.9 (30.9–34.8)	24.4 (22.8–26.2)	45.4 (42.1–48.6)	37.3 (33.0–41.9)	18.5 (17.6–19.4)	24.0 (22.7–25.3)	16.1 (13.6–18.4)
Baichuan M1 14B Instruct	41.7 (38.4–45.2)	28.6 (25.4–31.7)	32.4 (30.3–34.5)	57.6 (51.7–63.7)	41.0 (35.4–46.9)	41.5 (36.0–47.1)	29.0 (24.5–33.9)	38.8 (37.1–40.4)	42.2 (40.4–44.0)	57.1 (53.6–60.2)	38.8 (34.6–43.4)	19.1 (18.2–20.0)	**29.0 (27.6–30.4)*****	28.4 (25.7–31.4)
HuatuoGPT O1 8B	36.4 (32.8–39.8)	19.8 (17.3–22.2)	30.1 (28.4–31.9)	45.8 (39.5–52.2)	42.1 (36.6–47.8)	39.8 (34.8–44.5)	28.3 (22.9–34.0)	3.6 (2.8–4.4)	5.0 (4.0–6.1)	25.1 (22.4–27.7)	20.1 (16.6–23.9)	10.1 (9.4–10.8)	22.2 (20.5–23.7)	11.9 (10.1–14.0)
Meditron 3 8B	31.4 (26.9–35.8)	16.6 (13.5–19.6)	28.8 (27.0–30.5)	52.6 (46.0–59.2)	41.7 (36.7–46.6)	34.9 (29.2–40.9)	23.5 (18.3–29.0)	29.9 (28.2–31.4)	33.8 (32.1–35.8)	41.4 (38.0–44.4)	27.0 (23.0–30.7)	8.1 (7.3–8.8)	20.3 (18.6–22.1)	6.0 (4.6–7.6)
MedGemma 4B	27.9 (24.3–31.5)	14.0 (11.5–16.6)	22.7 (21.0–24.7)	46.8 (40.5–53.2)	39.7 (32.7–46.7)	39.3 (34.2–44.7)	29.7 (24.5–35.1)	36.5 (34.8–38.4)	23.7 (22.5–25.0)	39.6 (36.4–43.1)	25.7 (22.0–29.7)	7.5 (6.9–8.1)	19.3 (17.4–20.9)	3.6 (2.5–4.9)
GPT-OSS 120B	41.1 (38.3–44.1)	35.2 (32.4–37.9)	43.9 (41.7–46.0)	74.3 (68.8–79.5)	57.1 (51.0–63.5)	52.1 (46.3–57.7)	35.9 (30.7–41.3)	**54.7 (52.8–56.4)*****	23.0 (21.9–24.1)	**67.1 (63.5–70.2)**	43.4 (39.0–47.9)	**36.6 (35.3–37.8)*****	17.8 (17.1–18.6)	43.0 (39.8–46.0)
Llama 3.3 70B Instruct	40.0 (36.9–43.2)	27.5 (25.0–29.8)	25.2 (23.8–26.6)	59.9 (53.4–65.6)	50.7 (44.7–56.5)	47.5 (41.2–54.3)	37.5 (31.9–43.2)	42.7 (40.9–44.5)	31.8 (30.5–33.2)	61.7 (58.1–65.0)	44.2 (39.6–48.8)	22.5 (21.5–23.5)	24.7 (23.5–25.8)	33.0 (30.0–36.0)
Gemma 4 31B	44.2 (41.1–47.7)	33.1 (30.3–35.9)	44.4 (42.2–46.7)	**77.6 (72.6–82.6)**	56.7 (51.0–62.4)	50.8 (45.1–56.4)	34.7 (29.0–40.2)	41.1 (39.4–42.8)	33.9 (32.4–35.4)	61.8 (58.4–65.0)	39.4 (35.3–44.0)	23.6 (22.7–24.5)	25.9 (24.9–27.0)	45.3 (42.0–48.2)
Gemma 3 27B	42.1 (39.1–45.3)	29.9 (27.1–32.5)	37.2 (35.2–39.0)	60.8 (54.6–67.1)	58.4 (53.0–63.6)	47.2 (41.5–53.4)	36.6 (31.3–41.8)	43.5 (41.8–45.2)	26.5 (25.3–27.9)	50.9 (47.5–54.0)	32.4 (28.2–37.1)	17.9 (17.0–18.7)	24.3 (23.0–25.6)	26.1 (23.4–29.2)
GPT-OSS 20B	39.7 (36.5–42.9)	30.5 (28.1–33.2)	38.4 (35.4–40.8)	67.5 (61.9–73.4)	54.3 (48.9–60.0)	50.9 (45.2–56.5)	29.2 (24.7–33.6)	48.1 (46.1–49.9)	23.4 (22.3–24.7)	56.8 (53.6–60.2)	34.2 (29.9–38.8)	25.9 (24.8–26.9)	22.8 (21.6–23.9)	16.2 (13.8–18.6)
Gemma 4 E4B	39.4 (36.3–42.5)	27.2 (24.3–30.0)	32.2 (30.6–33.8)	56.1 (49.5–62.3)	53.9 (48.2–59.8)	50.1 (44.7–55.5)	31.6 (26.7–36.9)	39.6 (37.8–41.4)	31.7 (30.1–33.4)	53.7 (50.2–57.2)	39.8 (35.5–44.4)	24.2 (23.3–25.3)	26.6 (25.4–27.7)	18.1 (15.7–20.6)
Llama 3.1 8B Instruct	39.4 (36.2–42.7)	24.6 (22.2–27.0)	25.8 (24.3–27.3)	45.3 (38.4–51.9)	44.8 (39.5–50.2)	37.5 (32.6–42.7)	25.7 (20.7–30.7)	39.0 (37.3–40.7)	32.2 (30.8–33.6)	47.8 (44.5–50.9)	36.5 (32.4–41.1)	21.9 (21.0–22.9)	24.6 (23.4–25.7)	17.5 (15.1–19.9)

Values are mean percentages with 95% confidence intervals in parentheses across a 1,000-iteration bootstrap with replacement. For each metric, the highest-scoring Clin-REACT model and highest-scoring non-Clin-REACT model are selected. The higher of those two scores is shown in bold, and the lower score is underlined. Significance symbols indicate the paired Wilcoxon signed-rank p-value comparing those two models: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Abbreviations: Parent F1, parent-variable F1 score; Variable F1, individual-variable F1 score; Reasoning, reasoning score; SCT, SCT score; Decision, decision factors; Assess. Rec., assessment recall; Assess. Prec., assessment precision; Final Dx, final diagnosis accuracy. Included model groups: Clin-REACT, Open-source medical LLMs, Open-source general-purpose LLMs.

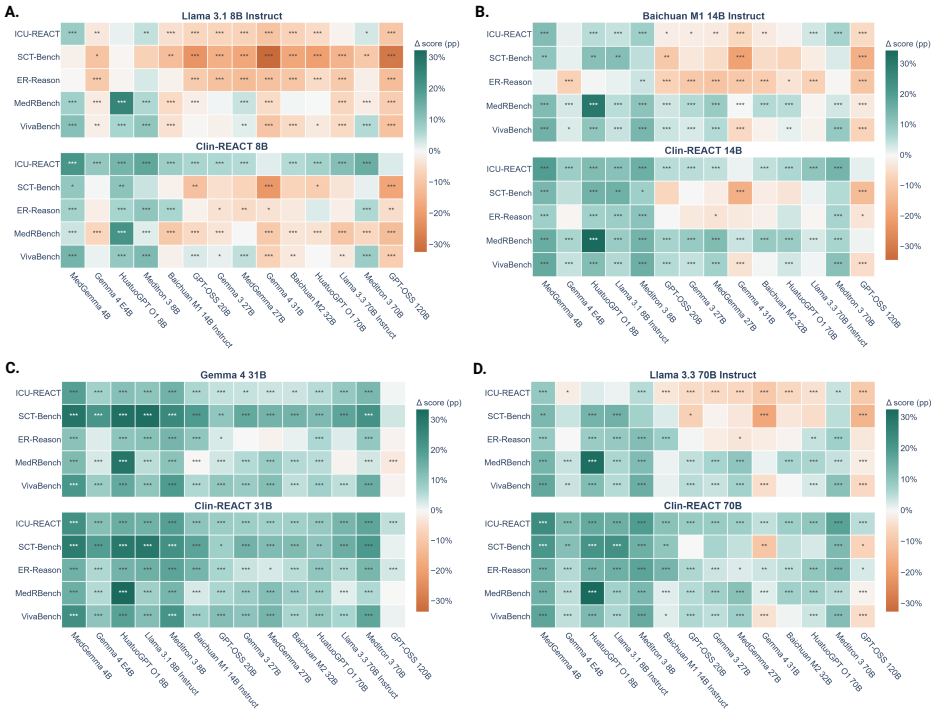


Fig. S6 Per-benchmark score differences of each Clin-REACT model and its backbone against baseline models across five clinical-reasoning benchmarks. Each panel pairs a backbone (top heatmap) with its fine-tuned Clin-REACT counterpart (bottom heatmap) at a single scale: (A) Llama 3.1 8B Instruct and Clin-REACT 8B, (B) Baichuan M1 14B Instruct and Clin-REACT 14B, (C) Gemma 4 31B and Clin-REACT 31B, and (D) Llama 3.3 70B Instruct and Clin-REACT 70B. Each cell reports the difference in score (Δ, percentage points) between the featured model (backbone or Clin-REACT) and a baseline model on one benchmark (rows), computed as the featured model minus the baseline using each benchmark's average-across-metrics score; green indicates the featured model scored higher, orange indicates the baseline scored higher, and color intensity represents the magnitude of the difference. Columns are the baseline models (all evaluated baselines except the panel's backbone), ordered by parameter count; rows are the five clinical-reasoning benchmarks (ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench), and the top and bottom heatmaps of each panel share the same columns and rows. Asterisks denote the statistical significance of the paired difference (two-sided Wilcoxon signed-rank test; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$). Each panel uses an independent color scale.

MedRBench, respectively, with a small, although non-significant, improvement of +0.8% on VivaBench ($p < 0.05$).

Among similarly sized models, Clin-REACT 8B demonstrated substantial advantages over other small baseline models. Compared with MedGemma 4B, HuatuoGPT O1 8B, and Meditron 3 8B, Clin-REACT 8B showed significant advantages across all five benchmarks, including advantages of +20.9%, +13.6%, and +16.8% on ICU-REACT, respectively (all $p < 0.001$). Its strongest relative gains were observed against HuatuoGPT O1 8B on MedRBench (+21.5%, $p < 0.001$) and against Meditron 3 8B

on ICU-REACT (+16.8%, $p < 0.001$), ER-Reason (+10.1%, $p < 0.001$), and VivaBench (+12.6%, $p < 0.001$). Clin-REACT 8B also outperformed Gemma 4 E4B on ICU-REACT (+9.5%, $p < 0.001$) and SCT-Bench (+0.2%, $p > 0.05$), although its performance was more mixed on ER-Reason and MedRBench.

Clin-REACT 14B similarly outperformed most comparably sized models, including the 8B baselines and GPT-OSS 20B. Relative to HuatuoGPT O1 8B, it achieved gains of +12.9%, +15.1%, +7.9%, +34.4%, and +12.6% across ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench, respectively (all $p < 0.001$). It also consistently exceeded Meditron 3 8B, with improvements ranging from +8.3% on SCT-Bench to +16.0% on ICU-REACT (all $p < 0.001$). Compared with GPT-OSS 20B, Clin-REACT 14B achieved significant gains on ICU-REACT (+5.5%, $p < 0.001$), MedRBench (+7.2%, $p < 0.001$), and VivaBench (+5.7%, $p < 0.001$), while differences on SCT-Bench and ER-Reason were not statistically significant.

Within the 20–32B parameter range, Clin-REACT 31B showed broad and consistent advantages over GPT-OSS 20B, Gemma 3 27B, MedGemma 27B, Gemma 4 31B, and Baichuan M2 32B. In particular, Clin-REACT 31B outperformed GPT-OSS 20B by +8.6% on ICU-REACT, +8.0% on SCT-Bench, +6.5% on ER-Reason, +6.3% on MedRBench, and +11.9% on VivaBench (all $p < 0.001$). It also achieved significant gains over Gemma 3 27B of +8.4%, +14.8%, +3.9%, +8.6%, and +10.7% across the five benchmarks, respectively. Compared with Gemma 4 31B, Clin-REACT 31B improved performance on ICU-REACT (+4.2%, $p < 0.001$), ER-Reason (+3.9%, $p < 0.001$), MedRBench (+2.9%, $p < 0.001$), and VivaBench (+1.9%, $p < 0.05$), with a non-significant decrease on SCT-Bench (-2.1%).

Among models in the 70B parameter range, Clin-REACT 70B consistently outperformed HuatuoGPT O1 70B, Llama 3.3 70B Instruct, and Meditron 3 70B. Compared with HuatuoGPT O1 70B, it achieved gains of +9.8% on ICU-REACT, +2.7% on SCT-Bench, +9.7% on ER-Reason, +7.2% on MedRBench, and +3.5% on VivaBench (all $p < 0.001$). Relative to Meditron 3 70B, Clin-REACT 70B showed particularly large improvements on ICU-REACT (+18.1%, $p < 0.001$), SCT-Bench (+12.5%, $p < 0.001$), ER-Reason (+13.2%, $p < 0.001$), and VivaBench (+11.8%, $p < 0.001$). These findings indicate that Clin-REACT fine-tuning conferred performance improvements not only over smaller models, but also over comparably sized general-purpose and medical LLMs. A visualization of the improvement of Clin-REACT models over their backbones when compared to baseline models can be seen in Fig. S6. Detailed results across all metrics can be found on Table S2.

S5 Ablations

We conducted two complementary ablation studies to characterize how Clin-REACT's performance depends on (i) the composition of self-supervised training tasks used to construct the ICU-REACT train datasets, and (ii) the scale of the training sets. Both studies were run on Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct backbones under identical pre-processing, split, and evaluation protocols, with all models evaluated on the held-out ICU-REACT test set.

S5.1 Training tasks

We tested generation of the ICU-REACT train datasets through five self-supervised augmentation tasks, each contributing a distinct form of clinical supervision. Each task was framed from the perspective of an ICU clinician and operated over patient context and decision-making question pairs:

- **Variable Selection Reasoning.** Given a patient context and a decision-making question, produced one coherent paragraph identifying which clinical variables are most relevant to the decision and explaining why, drawing on our curated ICU variable and taxonomy framework organized by category and sub-category. Rather than emitting a final answer, the task taught models to select and justify the decision-informing data that a clinician would review, spanning severity, trajectory, contraindications, response to therapy, and safety monitoring.
- **Reasoning Refinement.** Given a patient context, decision question, and an initial reasoning draft, produced a concise explanation of how the reasoning should be improved followed by a single revised reasoning paragraph. This critique-and-improve supervision targeted what is missing, overemphasized, or clinically misprioritized in the draft, sharpening intermediate reasoning quality and clinical coherence.
- **Question Generation.** Given only a patient context, produced one specific ICU decision-making question that would guide immediate clinical action (e.g., regarding fluids, vasopressors, ventilation, antibiotics, sedation, or diagnostics), teaching models to identify the salient, actionable decision implied by a case.
- **Context Generation.** Given only a decision-making question, produced one realistic patient context in which that question would be clinically applicable, complete with enough ICU-relevant detail (organ support, physiology, trajectory, comorbidities, acute problem) to justify why the question would be asked. This would potentially improve models' grounding in plausible clinical scenarios.
- **Context/Question Refinement.** Given an under specified initial context–question pair, produced a shared explanation of why it is too under specified for ICU decision-making, followed by all clinically distinct scenario-specific formulations, each with a scenario label, clinical rationale, refined question, and refined context. This task acted as quality control, disambiguating vague inputs into coherent, decision-ready scenarios.

S5.2 Task ablation

To isolate the contribution of each augmentation task, we trained both backbones on different task combinations, keeping the dataset size fixed, and evaluated each configuration on the ICU-REACT test set (Fig. S7). For Llama-3.1-8B-Instruct, the best configuration combined Variable Selection Reasoning and Reasoning Refinement, achieving an average score of 42.4 (Fig. S7A). The top three configurations each retained the Variable Selection Reasoning task, while the top two also retained the Reasoning Refinement task. Performance bottomed out at 39.4 and 38.2 when the

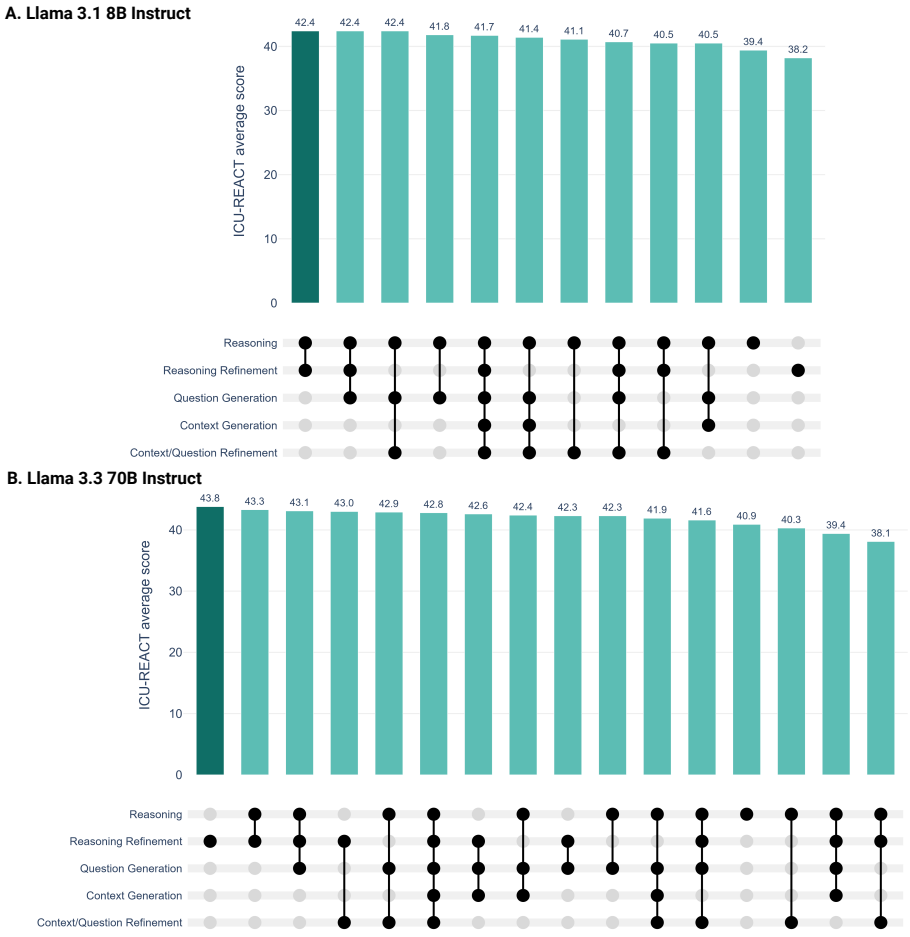


Fig. S7 Training task ablation. ICU-REACT average score for models trained on each combination of the five self-supervised augmentation tasks (Variable Selection Reasoning, Reasoning Refinement, Question Generation, Context Generation, and Context/Question Refinement) for (A) Llama-3.1-8B-Instruct and (B) Llama-3.3-70B-Instruct. Filled dots in the lower matrix denote the tasks included in a given configuration, connected vertically when combined; bars are sorted in descending order of average score, and the best-performing configuration in each panel is highlighted in dark teal.

model was trained on only Variable Selection Reasoning and Reasoning Refinement, respectively, suggesting that multi-task training benefited the 8B model the most. A slightly different pattern was observed for Llama-3.3-70B-Instruct, where the best configuration reached 43.8 by only training on the Reasoning Refinement task (Fig. S7B). Moreover, the top four configurations all included the Reasoning Refinement task. Performance generally declined with the addition of other tasks to the Reasoning Refinement task, suggesting that the 70B model benefited the most from a single task training on a critique task.

We decomposed performance into each ICU-REACT metric (parent variable F1, variable F1, and reasoning score) to assess whether the optimal task composition holds

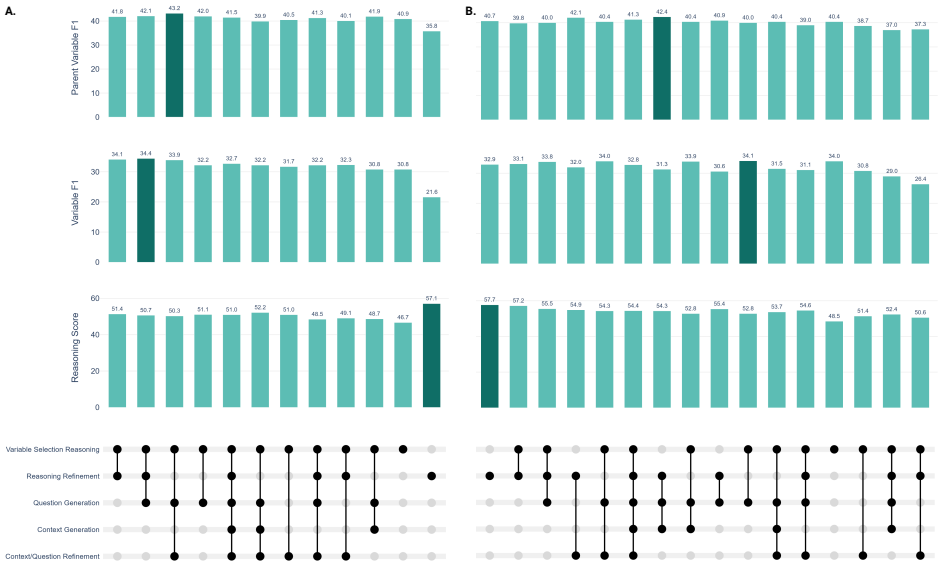


Fig. S8 Training task ablation decomposed by evaluation metric. ICU-REACT performance decomposed into parent variable F1 (top row), variable F1 (middle row), and reasoning score (bottom row) across training-task combinations for (A) Llama-3.1-8B-Instruct and (B) Llama-3.3-70B-Instruct. Filled dots in the lower matrix denote the tasks included in each configuration; within each metric row, the best-performing configuration is highlighted in dark teal.

across evaluation dimensions (Fig. S8). Across both backbones, Reasoning Refinement alone maximized the reasoning score (57.1 for 8B, 57.7 for 70B), but its effect on variable identification was scale-dependent. For Llama-3.1-8B-Instruct, training on Reasoning Refinement in isolation degraded both parent variable F1 (35.8) and variable F1 (21.6) to their lowest values, and pairing it with Variable Selection Reasoning was required to recover strong F1 performance (parent variable F1 41.8, variable F1 34.1) while retaining a still-robust reasoning score of 51.4 (Fig. S8A). For Llama-3.3-70B-Instruct, by contrast, Reasoning Refinement alone maintained robust variable identification (parent variable F1 40.7, variable F1 32.9) and achieved the highest reasoning score (57.7), remaining competitive with the best configurations which used Variable Selection Reasoning (parent variable F1 42.4, variable F1 34.1) (Fig. S8B). Thus, critique-based supervision consistently improved reasoning quality, but only the larger model absorbed this benefit without a corresponding trade-off in variable identification.

Taken together, these results indicate that the optimal training-task composition is backbone-dependent and metric-dependent, yet the Reasoning Refinement task emerged as the consistent throughline across both axes. Across both backbones, Reasoning Refinement drove the reasoning score, reflecting the central role of critique-based supervision in improving clinical reasoning quality. At 70B scale, this benefit came at no cost to variable identification, as Reasoning Refinement alone preserved robust parent variable and variable F1. At 8B scale, Reasoning Refinement in isolation degraded F1 performance, and pairing it with Variable Selection Reasoning was

necessary to recover strong variable identification while maintaining a robust reasoning score. In both cases, the Reasoning Refinement task was consistently present in the top-performing configurations, underscoring the central role of critique-based supervision in driving downstream ICU-REACT performance across model scales.

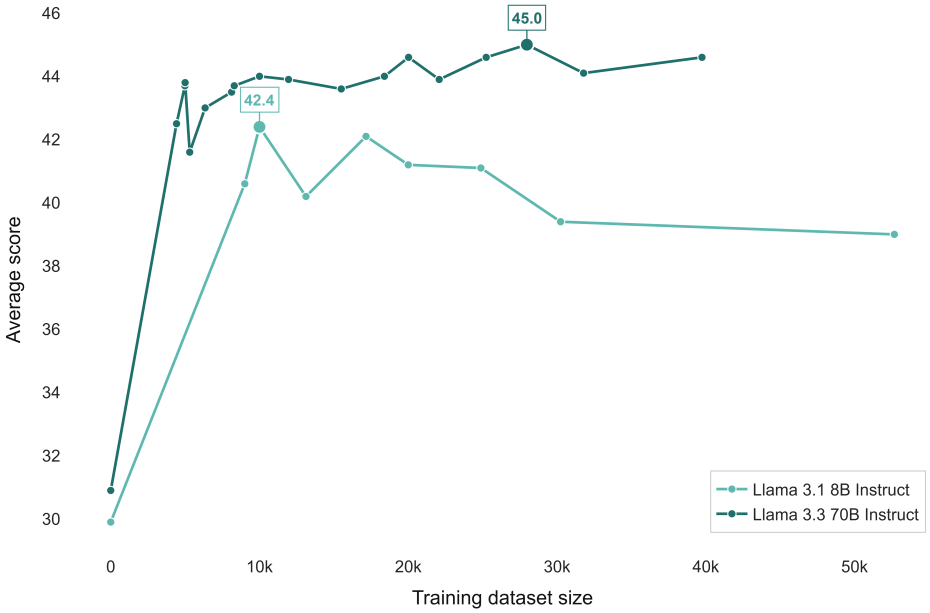


Fig. S9 Training dataset size ablation. Average ICU-REACT test score as a function of training dataset size for Llama-3.1-8B-Instruct (light teal) and Llama-3.3-70B-Instruct (dark teal). Each point is a model fine-tuned on the indicated number of examples; the leftmost point of each series (size 0) denotes the corresponding zero-shot instruct baseline. Highest performance for each model are highlighted with boxed scores.

S5.3 Dataset size ablation

We characterized scaling behavior by varying the number of training examples while holding the task composition fixed (Reasoning Refinement and Variable Selection Reasoning for Llama 3.1 8B Instruct and Reasoning Refinement for Llama 3.3 70B Instruct) (Figure S9). Both backbones exhibited steep early gains followed by rapid saturation. Starting from their zero-shot baselines (29.9 for the 8B model and 30.9 for the 70B model), both models improved sharply within the first several thousand examples. Llama-3.3-70B-Instruct rose above 43 by roughly 8,000 examples and continued to improve gradually thereafter, peaking at 45.0 near 28,000 examples before plateauing. Llama-3.1-8B-Instruct peaked earlier, reaching 42.4 at around 10,000 examples, after which performance was unstable and declined gradually with additional data, settling below 40 at the largest training sizes. This divergence suggests that the larger backbone continues to benefit from additional training data well beyond the point at

which the smaller model saturates, and that the 8B model is prone to mild degradation when trained on substantially more data than its effective capacity supports. Taken together, the two ablations indicate that a compact, reasoning-focused training mixture on the order of 10k–30k examples captures most of the attainable performance, with the optimal scale increasing with backbone size.

S6 Comparison with Frontier Models

We benchmarked the four Clin-REACT models against four proprietary frontier LLMs across five clinical-reasoning benchmarks: GPT-5.2, GPT-5 Mini, Claude 4.6 Sonnet, and Gemini 3.1 Pro (Fig. S10; Table S3). Averaged across benchmarks, Clin-REACT 31B achieved a macro clinical-reasoning score of 50.4% (SD, 15.5), closely matching GPT-5 Mini at 50.5% (SD, 13.3) and approaching Gemini 3.1 Pro at 51.4% (SD, 12.9), GPT-5.2 at 52.9% (SD, 15.1), and Claude 4.6 Sonnet at 53.2% (SD, 13.2). This performance was achieved despite Clin-REACT 31B being an open-weight model roughly an order of magnitude smaller and incurring no per-token inference cost, whereas frontier output prices ranged from $2 to $15 per 1M tokens (Fig. S10A).

On three of the five benchmarks (ICU-REACT, SCT-Bench, and ER-Reason) the best Clin-REACT model was statistically indistinguishable from all frontier models (Fig. S10B). On ICU-REACT, Clin-REACT 70B achieved an aggregate score of 45.0% (95% CI, 42.8–47.4), compared with 47.9% (45.8–49.9) for GPT-5.2, 45.9% (44.1–47.9) for Claude 4.6 Sonnet, 44.2% (42.1–46.2) for Gemini 3.1 Pro, and 44.0% (41.7–46.3) for GPT-5 Mini. At the component level, Clin-REACT 31B achieved the highest parent-variable F1 of any model (45.1% versus 43.9% for Gemini 3.1 Pro), and its remaining sub-metrics did not differ significantly from the best frontier scores (Table S3). On SCT-Bench, Clin-REACT 31B scored 75.5% (69.8–81.0), equal to Claude 4.6 Sonnet at 75.5% (69.8–80.7) and close to GPT-5.2 at 77.5% (72.0–82.2). On ER-Reason, Clin-REACT 31B scored 51.4% (46.5–55.8), compared with 52.8% (48.2–57.5) for GPT-5.2, 52.2% (47.8–56.7) for Gemini 3.1 Pro, and 51.4% (46.9–55.5) for Claude 4.6 Sonnet. Only the ER-Reason differential-diagnosis sub-metric reached significance (GPT-5.2, 58.3%, versus Clin-REACT 31B, 54.2%; $p < 0.05$) (Table S3). Thus, on focused reasoning and information-retrieval tasks, the Clin-REACT models were competitive with systems many times their size and cost.

Frontier models retained a clearer advantage on MedRBench and VivaBench. On MedRBench, the highest-performing Clin-REACT model, Clin-REACT 14B, achieved 47.9% (47.0–48.7), compared with 51.8% (50.9–52.6) for Claude 4.6 Sonnet, 50.2% (49.4–51.1) for GPT-5.2, 50.2% (49.4–51.1) for Gemini 3.1 Pro, and 49.8% (49.0–50.5) for GPT-5 Mini; all four frontier models significantly exceeded the best Clin-REACT model ($p < 0.001$). On VivaBench, Clin-REACT 31B achieved 33.5% (32.2–34.8), compared with 41.4% (40.1–42.7) for Claude 4.6 Sonnet, 38.2% (36.8–39.4) for Gemini 3.1 Pro, 37.3% (35.9–38.6) for GPT-5 Mini, and 36.3% (35.0–37.6) for GPT-5.2. Three of the four frontier models significantly exceeded the best Clin-REACT model on VivaBench, whereas the difference from GPT-5.2 was not significant.

At the component level, frontier models showed their largest advantages on several MedRBench and VivaBench metrics. On MedRBench, GPT-5 Mini achieved higher

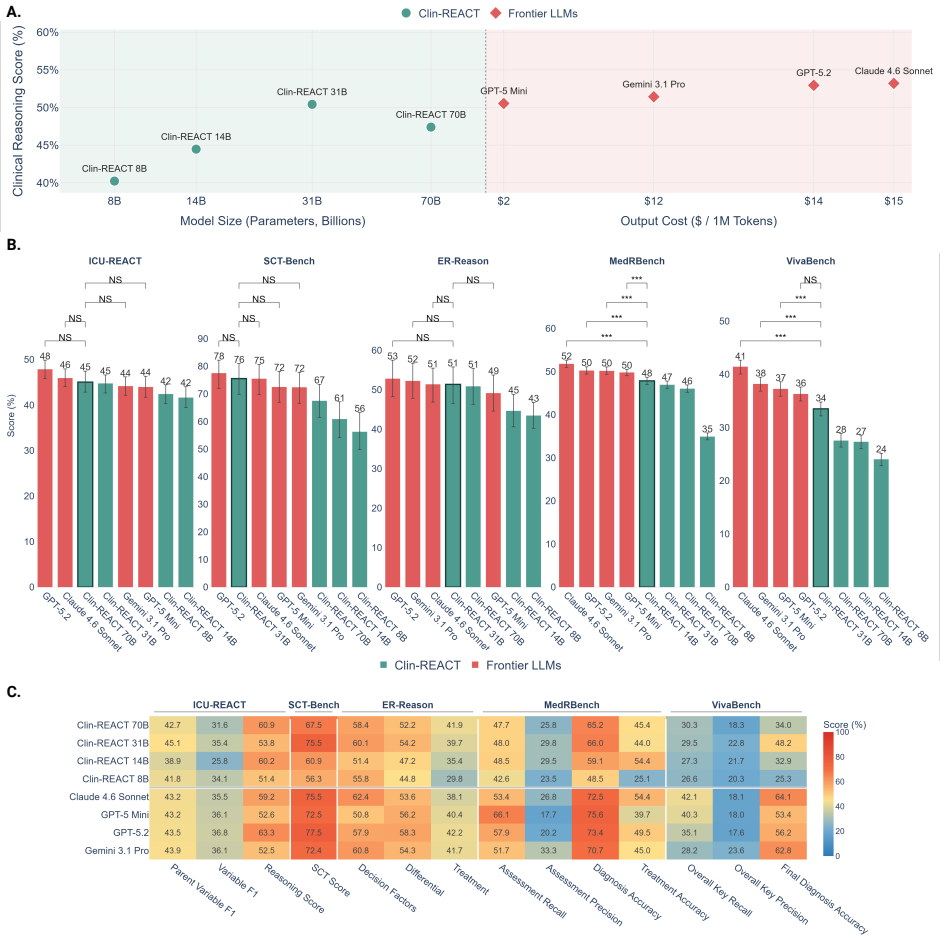


Fig. S10 Comparison of Clin-REACT models against proprietary frontier LLMs on clinical reasoning. (A) Clinical-reasoning score across five benchmarks for both model classes, plotted against parameter count (left region, in billions) for Clin-REACT models (circles, $n = 4$) and against output cost (right region, $ per 1M tokens) for frontier LLMs (diamonds, $n = 4$), with a vertical divider separating the two scales. Each clinical-reasoning score is the mean across five clinical-reasoning benchmarks (ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench). (B) Per-benchmark comparison of the four Clin-REACT models (teal) against the four frontier LLMs (red) across the five clinical-reasoning benchmarks. Bars are ordered by score within each panel, and the highest-scoring Clin-REACT model per benchmark is outlined. Error bars denote 95% CIs; brackets indicate pairwise comparisons between the best Clin-REACT model on each benchmark and frontier LLMs (two-sided Wilcoxon signed-rank test; NS, not significant; ***, $p < 0.001$). (C) Heatmap of component sub-metrics underlying each aggregate benchmark score in (B), grouped by benchmark: ICU-REACT (Parent Variable F1, Variable F1, Reasoning Score), SCT-Bench (SCT Score), ER-Reason (Decision Factors, Differential, Treatment), MedRBench (Assessment Recall, Assessment Precision, Diagnosis Accuracy, Treatment Accuracy), and VivaBench (Overall Key Recall, Overall Key Precision, Final Diagnosis Accuracy). Cell color encodes score (0–100%), with warmer colors indicating higher scores.

Table S3 Performance of Clin-REACT models against proprietary frontier models across clinical reasoning benchmarks.

Model	ICU-REACT			SCT-Bench	ER-Reason			MedR-Bench				VivaBench		
	Parent F1	Variable F1	Reasoning	SCT	Decision	Differential	Treatment	Assess. Rec.	Assess. Prec.	Diagnosis	Treatment	Overall Recall	Overall Precision	Final Dx
Clin-REACT 70B	42.7 (39.4–46.3)	31.6 (28.9–34.3)	60.9 (58.2–63.4)	67.5 (61.6–73.0)	58.4 (52.9–63.5)	52.2 (46.2–57.9)	41.9 (36.3–47.3)	47.7 (46.1–49.6)	25.8 (24.7–27.1)	65.2 (62.1–68.3)	45.4 (41.1–49.8)	30.3 (29.1–31.4)	18.3 (17.6–19.0)	34.0 (31.2–37.0)
Clin-REACT 31B	**45.1 (42.0–48.3)**	35.4 (32.9–37.9)	53.8 (51.7–55.8)	75.5 (69.8–80.6)	60.1 (54.2–65.7)	54.2 (48.3–60.7)	39.7 (34.1–45.3)	48.0 (46.2–49.7)	29.8 (28.4–31.1)	66.0 (62.7–69.0)	44.0 (39.8–48.3)	29.5 (28.4–30.6)	22.8 (21.9–23.7)	48.2 (45.4–51.5)
Clin-REACT 14B	38.9 (35.5–42.6)	25.8 (23.2–28.3)	60.2 (57.8–62.8)	60.9 (54.7–66.8)	51.4 (45.5–57.3)	47.2 (42.1–52.7)	35.4 (30.1–40.9)	48.5 (46.9–50.3)	29.5 (28.2–30.6)	59.1 (55.7–62.2)	54.4 (50.0–58.5)	27.3 (26.3–28.4)	21.7 (20.8–22.7)	32.9 (29.7–35.9)
Clin-REACT 8B	41.8 (38.7–44.7)	34.1 (31.5–36.8)	51.4 (49.1–53.6)	56.3 (49.7–62.2)	55.8 (50.9–60.6)	44.8 (39.9–49.7)	29.8 (25.4–34.5)	42.6 (41.0–44.3)	23.5 (22.3–24.8)	48.5 (45.1–52.0)	25.1 (21.6–29.3)	26.6 (25.4–27.6)	20.3 (19.5–21.1)	25.3 (22.3–28.2)
Claude 4.6 Sonnet	43.2 (40.4–46.2)	35.5 (33.4–37.6)	59.2 (56.8–61.5)	75.5 (70.1–80.8)	**62.4 (57.1–67.6)**	53.6 (48.0–59.4)	38.1 (32.3–43.7)	53.4 (51.6–55.2)	26.8 (25.5–28.2)	72.5 (69.5–75.6)	**54.4 (50.0–59.1)**	**42.1 (40.9–43.4)*****	18.1 (17.4–18.7)	**64.1 (61.0–67.1)*****
GPT-5 Mini	43.2 (40.0–46.6)	36.1 (33.6–38.8)	52.6 (50.4–55.0)	72.5 (66.9–77.7)	50.8 (44.1–57.0)	56.2 (50.8–61.9)	40.4 (34.7–45.9)	**66.1 (64.4–67.7)*****	17.7 (17.0–18.6)	**75.6 (72.8–78.5)*****	39.7 (34.9–44.5)	40.3 (39.1–41.6)	18.0 (17.3–18.7)	53.4 (50.5–56.5)
GPT-5.2	43.5 (40.3–46.7)	**36.8 (34.2–39.2)**	**63.3 (61.0–65.5)**	**77.5 (72.4–82.4)**	57.9 (50.9–63.9)	**58.3 (52.9–63.8)***	**42.2 (36.8–47.2)**	57.9 (55.9–59.7)	20.2 (19.3–21.1)	73.4 (70.3–76.2)	49.5 (44.8–54.3)	35.1 (33.9–36.3)	17.6 (16.9–18.3)	56.2 (53.2–59.2)
Gemini 3.1 Pro	43.9 (40.8–47.2)	36.1 (33.6–38.5)	52.5 (50.3–54.6)	72.4 (66.7–77.6)	60.8 (54.9–66.1)	54.3 (48.7–59.9)	41.7 (35.5–46.8)	51.7 (49.9–53.5)	**33.3 (31.7–34.8)*****	70.7 (67.5–73.6)	45.0 (40.5–49.2)	28.2 (27.1–29.3)	**23.6 (22.6–24.6)**	62.8 (59.6–66.0)

Values are mean percentages with 95% confidence intervals in parentheses across a 1,000-iteration bootstrap with replacement. For each metric, the highest-scoring Clin-REACT model and highest-scoring frontier model are selected. The higher of those two scores is shown in bold, and the lower score is underlined. Significance symbols indicate the paired Wilcoxon signed-rank p-value comparing those two models: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Abbreviations: Parent F1, parent-variable F1 score; Variable F1, individual-variable F1 score; Reasoning, reasoning score; SCT, SCT score; Decision, decision factors; Assess. Rec., assessment recall; Assess. Prec., assessment precision; Final Dx, final diagnosis accuracy. Included model groups: Clin-REACT, Frontier LLMs.

assessment recall than Clin-REACT 14B (66.1% versus 48.5%) and higher diagnosis accuracy than Clin-REACT 31B (75.6% versus 66.0%). However, Clin-REACT 14B matched Claude 4.6 Sonnet on treatment accuracy (54.4% for both models). On VivaBench, Claude 4.6 Sonnet achieved higher overall key recall than Clin-REACT 70B (42.1% versus 30.3%) and higher final-diagnosis accuracy than Clin-REACT 31B (64.1% versus 48.2%). In contrast, overall key precision was similar between Gemini 3.1 Pro and Clin-REACT 31B (23.6% versus 22.8%), with the difference not reaching statistical significance. Taken together, these results indicate that scale- and cost-efficient open models can be trained to closely approach or even match proprietary frontier models on targeted clinical-reasoning tasks.

S7 Performance stratified by clinical content

Clin-REACT models demonstrated gains over their corresponding backbones across clinical categories, with several related content areas showing parallel improvements across benchmarks (Fig. S11). For example, Clin-REACT 14B improved substantially in ICU-REACT Nutrition and Metabolic Support (+8.4 percentage points [pp]) and in the related MedRBench Metabolic Problems category (+12.3 pp), while also gaining on VivaBench Cardiovascular and Metabolic conditions (+3.0 pp). Similarly, Clin-REACT 31B improved on ICU-REACT Sepsis and Severe Infections (+8.1 pp), MedRBench Infections (+3.0 pp), and VivaBench Infectious Disease and Immunology (+4.6 pp). Hematologic content also showed cross-benchmark gains for several variants, including Clin-REACT 70B on ICU-REACT Hematologic/Coagulation Issues (+13.4 pp) and VivaBench Hematology/Oncology/Other (+7.8 pp). These patterns suggest that training gains in some ICU-REACT topics transferred to clinically related categories in external benchmarks, although the correspondence was not uniform across model sizes or datasets. For example, Clin-REACT 8B improved across all ICU-REACT topics but regressed across all MedRBench disorder groups, while Clin-REACT 70B showed strong ICU-REACT gains but declined in selected VivaBench specialties, most notably Neurological/Psychiatric conditions (−7.1 pp).

Compared with the evaluated open-source baseline models, Clin-REACT models ranked strongly on most clinical categories across benchmarks (Fig. S12). A Clin-REACT variant achieved the highest score in each of the seven ICU-REACT topic categories, and multiple Clin-REACT variants frequently occupied the leading positions within the same category. Clin-REACT models also achieved the highest category-level scores in four of the six MedRBench disorder groups, including Cancers, Infections, Metabolic Problems, and Pregnancy and Reproduction. On VivaBench, Clin-REACT models led or tied for the highest score in most specialty groups, including Infectious Disease and Immunology, Cardiovascular and Metabolic, Endocrine and Reproductive, Neurological/Psychiatric, Hematology/Oncology/Other, and Respiratory conditions. Performance was comparatively weaker in the Pediatric and Gastrointestinal specialties, where general-purpose baseline models ranked highest. Detailed metric-level differences are provided in Fig. S13.

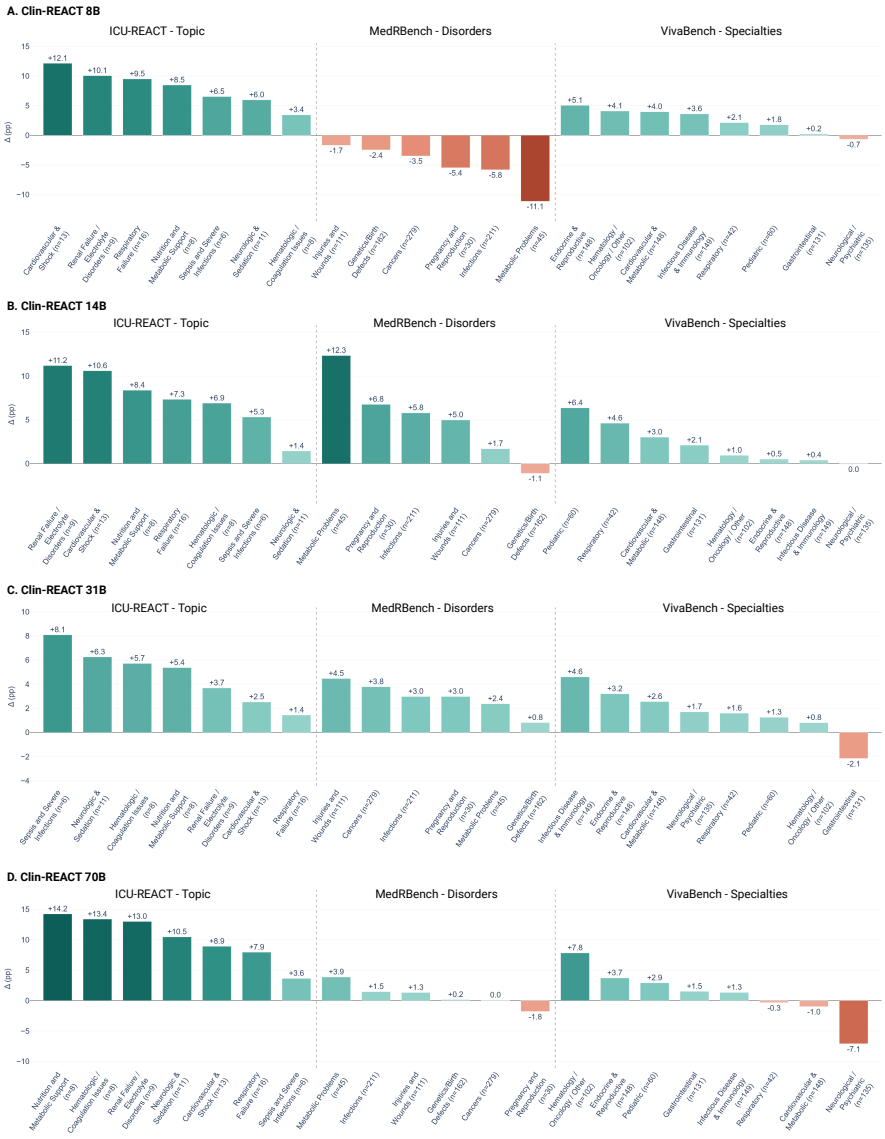


Fig. S11 Per-clinical content training gains of Clin-REACT models over backbones across clinical-reasoning benchmarks. Gains are shown for (A) Clin-REACT 8B (backbone: Llama 3.1 8B Instruct), (B) Clin-REACT 14B (backbone: Baichuan M1 14B Instruct), (C) Clin-REACT 31B (backbone: Gemma 4 31B), and (D) Clin-REACT 70B (backbone: Llama 3.3 70B Instruct). Each bar reports the change in score (Δ, percentage points) for a single clinical category, computed as the Clin-REACT model minus its corresponding backbone using each benchmark's average-across-metrics score; bar color encodes the sign and magnitude of the change, with teal representing improvement over the backbone, red representing regression, and darker shades denoting larger magnitudes. Within each benchmark section, bars are sorted in descending order of Δ, so category order varies across panels; numeric labels show each category's Δ and axis labels show the number of test samples (n). The three sections per panel, separated by dashed vertical lines, correspond to ICU-REACT topics (7 categories), MedRBench disorder groups (6 categories), and VivaBench specialty groups (8 categories).

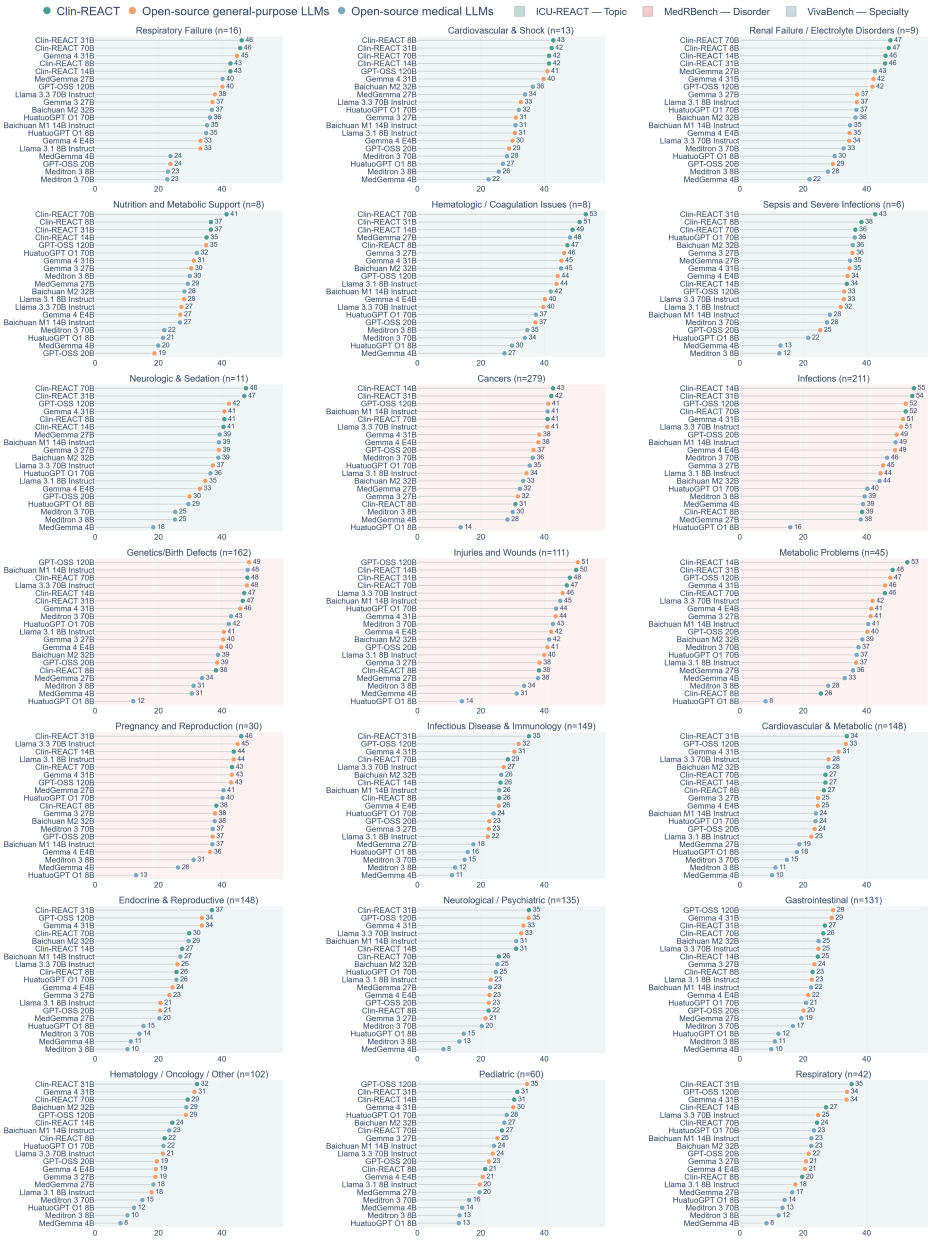


Fig. S12 Per-clinical content comparison of Clin-REACT and baseline models across clinical-reasoning benchmarks. Each panel ranks all evaluated models by score within a single clinical category, drawn from one of the three benchmarks that carry per-sample categorical labels: ICU-REACT topics (green panels, 7 categories), MedRBench disorder groups (pink panels, 6 categories), and VivaBench specialty groups (blue panels, 8 categories). Panel titles give the category name and its number of test samples (n). Within each panel, points are ordered by score (highest at top, ranked independently per panel) and colored by model class: Clin-REACT (teal, $n = 4$), open-source general purpose LLMs (orange, $n = 9$), and open-source medical LLMs (blue, $n = 6$); numeric labels give each model's score within the category. Scores are calculated as the average score across metrics within each benchmark.

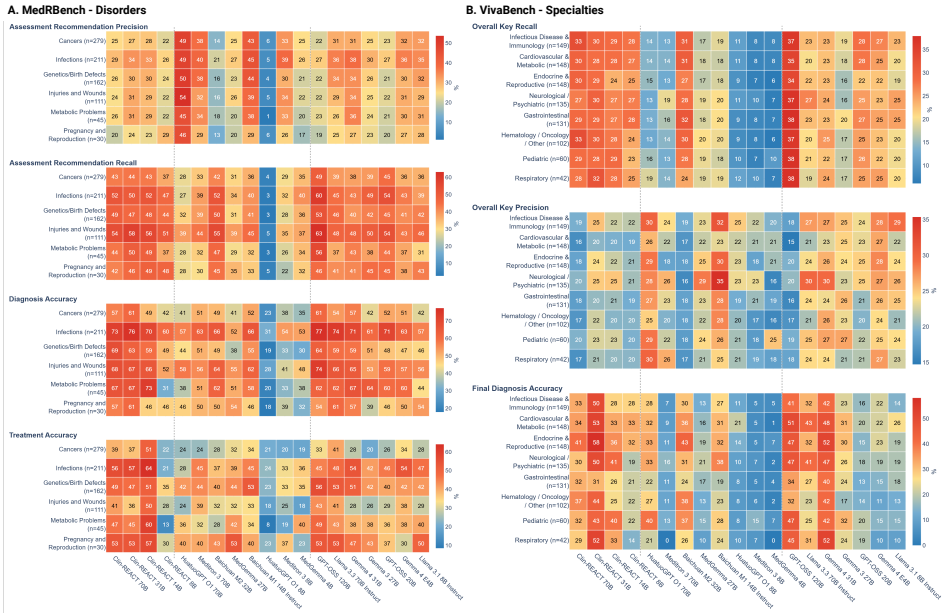


Fig. S13 Per-clinical content, per-metric heatmap of Clin-REACT and baseline models' performance on MedRBench and VivaBench. Each cell reports a model's score (%) within a single clinical category (rows) for one evaluation metric (heatmap blocks). Columns comprise all evaluated models, grouped by class and separated by dashed vertical lines: Clin-REACT (left, $n = 4$), open-source medical LLMs (center, $n = 8$), and open-source general purpose LLMs (right, $n = 7$). Row labels give the category name and its number of test samples (n); numeric labels give each cell's score. Warmer colors indicate higher scores, with each metric block using an independent color scale such shading is comparable within a block. (A) MedRBench disorder groups (6 categories) across four metrics: Assessment Recommendation Precision, Assessment Recommendation Recall, Diagnosis Accuracy, and Treatment Accuracy. (B) VivaBench specialty groups (8 categories) across three metrics: Overall Key Recall, Overall Key Precision, and Final Diagnosis Accuracy.

S8 Information seeking performance by data category

Clin-REACT training produced category-specific improvements in information retrieval, with the most consistent cross-benchmark transfer observed for imaging-related information (Fig. S14). All four Clin-REACT variants improved imaging F1 on both ICU-REACT and VivaBench. The largest gains were achieved by Clin-REACT 70B and Clin-REACT 8B on ICU-REACT imaging (+26.6 and +22.4 percentage points [pp], respectively), accompanied by gains of +7.8 and +6.6 pp on VivaBench imaging. Clin-REACT 14B similarly improved imaging retrieval on ICU-REACT (+8.3 pp) and VivaBench (+6.6 pp), while Clin-REACT 31B showed smaller but positive gains on both benchmarks (+0.6 and +1.6 pp). This consistent pattern suggests that the retrieval improvements learned from ICU-focused training transferred particularly well to imaging information in an external clinical-reasoning benchmark.

Other gains were more dependent on model size and data category. Clin-REACT 8B showed substantial ICU-REACT improvements for medications (+26.6 pp), physiology (+13.1 pp), laboratory measurements (+9.8 pp), and scores and assessments

(+8.7 pp), while also improving VivaBench history (+1.5 pp) and investigation retrieval (+4.4 pp). Clin-REACT 31B and 70B likewise improved ICU-REACT physiology and scores and assessments, with Clin-REACT 31B showing its largest VivaBench gain for history (+8.7 pp) and Clin-REACT 70B for investigation (+9.4 pp). Clin-REACT 14B exhibited a different pattern, with weaker or negative gains in several ICU-REACT categories but consistent improvements across all four VivaBench finding categories, reaching +7.4 pp for investigation. Thus, gains transferred across benchmarks for several related information types, but the magnitude and direction of transfer varied across model sizes. The corresponding category-level precision and recall profiles indicate that these differences reflected distinct retrieval tradeoffs across model families and data types; detailed values are provided in Fig. S15.

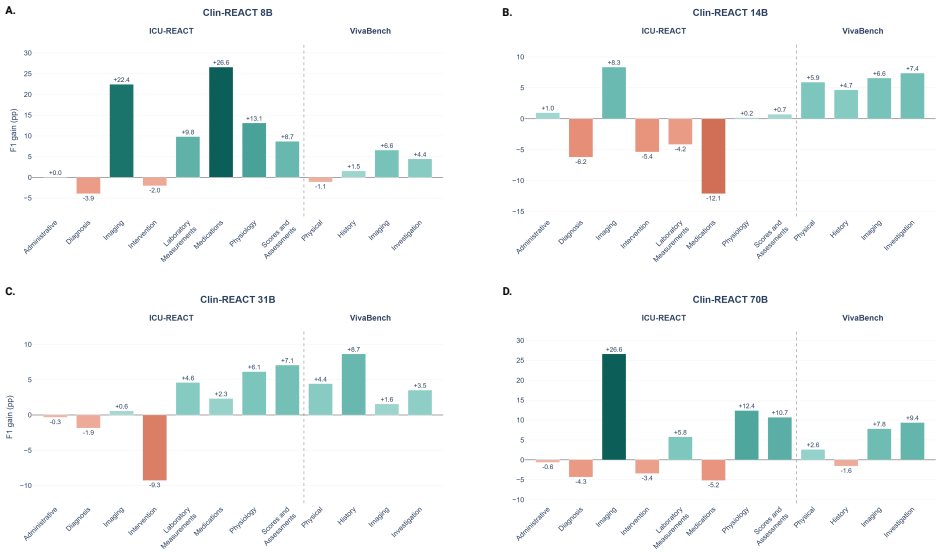


Fig. S14 Per-category information-retrieval gains of Clin-REACT models over backbones on clinical-reasoning benchmarks. Gains are shown for (A) Clin-REACT 8B (backbone: Llama 3.1 8B Instruct), (B) Clin-REACT 14B (backbone: Baichuan M1 14B Instruct), (C) Clin-REACT 31B (backbone: Gemma 4 31B), and (D) Clin-REACT 70B (backbone: Llama 3.3 70B Instruct) across ICU-REACT and VivaBench. Each bar reports the change in information-retrieval F1 (Δ, percentage points) for a single content category, computed as the Clin-REACT model's F1 minus its corresponding backbone's F1; bar color encodes the sign and magnitude of the change, with teal representing improvement over the backbone, red representing regression, and darker shades denoting larger magnitudes. Categories appear in a fixed order that is consistent across panels, and numeric labels give each category's Δ. The two sections per panel, separated by a dashed vertical line, correspond to ICU-REACT variable categories (8 categories) and VivaBench finding categories (4 categories).

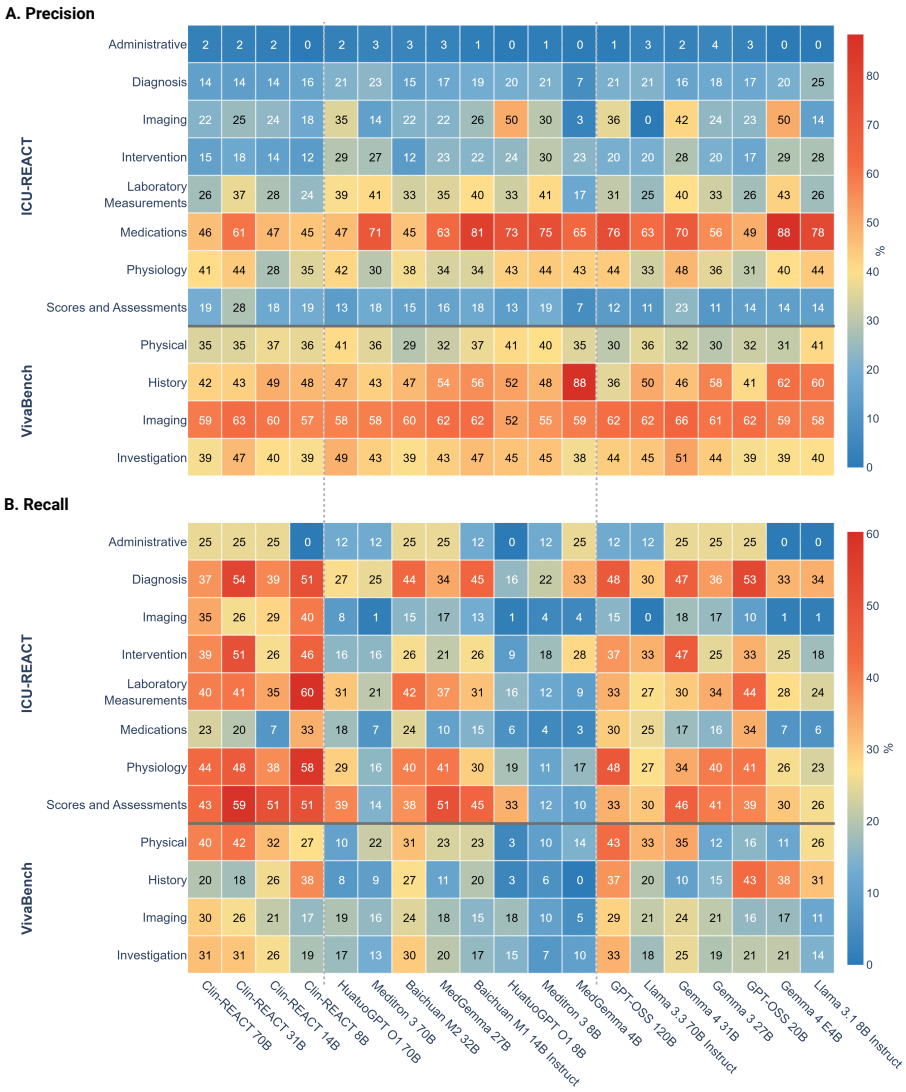


Fig. S15 Per-category heatmaps of information-retrieval of Clin-REACT and baseline models on ICU-REACT and VivaBench. Information-retrieval is shown as (A) precision and (B) recall. Each cell reports a model's value (%) for a single content category (rows). Columns comprise all evaluated models, grouped by class and separated by dashed vertical lines: Clin-REACT (left, $n = 4$), open-source medical LLMs (center, $n = 8$), and open-source general purpose LLMs (right, $n = 7$). Rows are grouped and separated by a horizontal line into ICU-REACT variable categories (8 categories) and VivaBench finding categories (4 categories); numeric labels give each cell's value. Warmer colors indicate higher values, with each panel using an independent color scale such that shading is comparable within a panel.

Table S4 Performance of Clin-REACT models against their backbones on general medical multiple-choice benchmarks.

Model	Average	MMLU-Med	MedMCQA	MedQA	MedXpertQA
Llama 3.1 8B Instruct	43.3	64.0	47.5	47.1	14.8
Clin-REACT 8B	34.6 ▼ (−8.7)	55.2 ▼ (−8.8)	42.0 ▼ (−5.5)	27.3 ▼ (−19.8)	13.8 ▼ (−1.0)
Baichuan M1 14B Instruct	56.5	81.4	62.5	61.2	21.0
Clin-REACT 14B	55.3 ▼ (−1.2)	78.8 ▼ (−2.6)	61.3 ▼ (−1.2)	59.5 ▼ (−1.7)	21.7 ▲ (+0.7)
Gemma-4-31B-It	80.2	94.2	78.5	92.4	55.8
Clin-REACT 31B	78.8 ▼ (−1.4)	94.9 ▲ (+0.7)	78.4 ▼ (−0.1)	89.8 ▼ (−2.6)	52.2 ▼ (−3.6)
Llama 3.3 70B Instruct	57.5	73.6	57.0	77.8	21.5
Clin-REACT 70B	56.3 ▼ (−1.2)	69.0 ▼ (−4.6)	57.2 ▲ (+0.2)	73.6 ▼ (−4.2)	25.3 ▲ (+3.8)

All values are accuracy (%). Each block pairs an instruction-tuned backbone (top) with its corresponding Clin-REACT model (bottom); arrows indicate the change relative to that backbone.

S9 Multiple choice medical benchmarks vs clinical reasoning

To assess whether Clin-REACT training preserved general medical knowledge, we compared each Clin-REACT model with its corresponding backbone across four medical multiple-choice benchmarks: MMLU-Med, MedMCQA, MedQA, and MedXpertQA. Clin-REACT training generally preserved performance on these benchmarks, particularly for the larger model variants (Table S4). Relative to their corresponding backbones, Clin-REACT 14B, 31B, and 70B showed only modest reductions in average multiple-choice accuracy of 1.2, 1.4, and 1.2 percentage points, respectively. These changes were not uniformly negative across individual benchmarks: Clin-REACT 31B improved on MMLU-Med (+0.7 points), Clin-REACT 70B improved on MedMCQA (+0.2) and MedXpertQA (+3.8), and Clin-REACT 14B improved on MedXpertQA (+0.7). Clin-REACT 8B showed a larger average reduction of 8.7 points, driven primarily by MedQA (−19.8 points), indicating that preservation of multiple-choice performance was less consistent at the smallest model scale.

To examine whether performance on medical multiple-choice benchmarks reflects broader clinical-reasoning ability, we conducted a correlation analysis between each model's average score across four multiple-choice benchmarks (MMLU-Med, MedMCQA, MedQA, and MedXpertQA) and its average score across five clinical-reasoning benchmarks (ICU-REACT, SCT-Bench, ER-Reason, MedRBench, and VivaBench) (Fig. S16). Among the baseline general-purpose and medical models, multiple-choice and clinical-reasoning performance were positively correlated ($R^2 = 0.67$; Spearman $\rho = 0.78$), but substantial departures from this overall trend indicated that strong multiple-choice performance did not consistently translate into strong clinical reasoning. For example, Meditron 3 70B, HuatuoGPT O1 70B, and MedGemma 27B achieved higher average multiple-choice scores than Llama 3.3 70B Instruct (61.4%, 66.8%, and 67.1% versus 57.5%, respectively), yet all three performed worse on the clinical-reasoning benchmarks (35.3%, 40.8%, and 40.6% versus 41.6%). Among the smaller models, the same pattern was evident when comparing Llama 3.1 8B Instruct

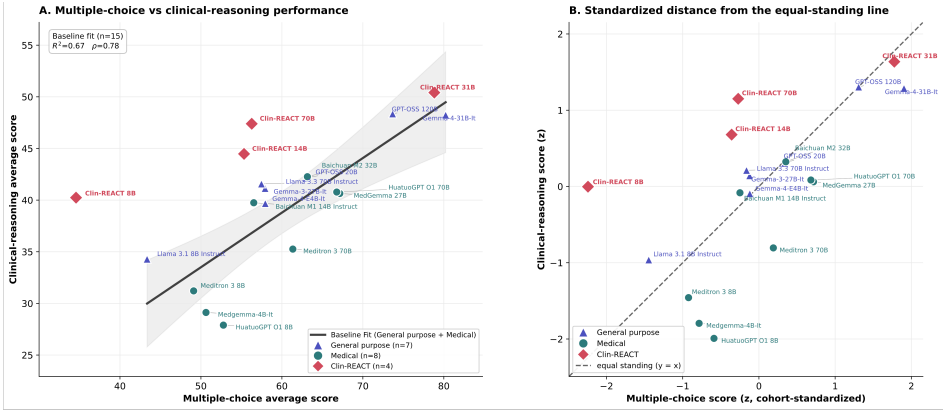


Fig. S16 Relationship between medical multiple-choice and clinical reasoning benchmarks performance. (A) Clinical-reasoning average score versus multiple-choice average score for all evaluated models. Each benchmark score is first computed by averaging its constituent sub-metrics, and each axis is then the mean across benchmarks (four multiple-choice benchmarks: MMLU-Med, MedMCQA, MedQA, MedXpertQA; five clinical-reasoning benchmarks: ER-Reason, ICU-REACT, MedRBench, SCT-Bench, VivaBench). Marker shape denotes model class: general-purpose (triangles, $n = 7$), open-source medical (circles, $n = 8$), and Clin-REACT (diamonds, $n = 4$). The solid line is an ordinary least-squares fit to the baseline models only (general-purpose and medical, $n = 15$; shaded band, 95% confidence interval); Clin-REACT models are overlaid but excluded from the fit so that the line describes the multiple-choice-to-reasoning relationship among existing models. Goodness of fit for the baseline relationship is inset ($R^2 = 0.67$, Spearman $\rho = 0.78$). (B) The same models after standardizing each axis to z-scores across the full cohort ($n = 19$). The dashed diagonal ($y = x$) marks equal relative standing on the two benchmark families: models above the line rank higher on clinical reasoning than on multiple choice, and models below the line rank lower.

with medical models of a similar scale. HuatuoGPT O1 8B, MedGemma 4B, and Meditron 3 8B all achieved higher average multiple-choice scores than Llama 3.1 8B Instruct (52.8%, 50.6%, and 49.1% versus 43.4%, respectively), yet each obtained a lower average clinical-reasoning score (27.9%, 29.1%, and 31.2% versus 34.3%). These examples show that models with stronger performance on knowledge-focused multiple-choice tasks may still underperform models with lower multiple-choice scores when evaluated on benchmarks requiring information retrieval, evidence integration, and clinical decision-making.

Clin-REACT models further illustrated this distinction. Clin-REACT 14B and Clin-REACT 70B achieved clinical-reasoning scores of 44.5% and 47.4%, respectively, despite multiple-choice scores of 55.3% and 56.3%; both therefore outperformed several models with substantially higher multiple-choice scores, including MedGemma 27B and HuatuoGPT O1 70B. The contrast was strongest for Clin-REACT 8B, which achieved the lowest multiple-choice average in the cohort (34.6%) but a clinical-reasoning score of 40.2%, exceeding Llama 3.1 8B Instruct (34.3%), Meditron 3 70B (35.3%), and several medical models with considerably higher multiple-choice performance. After cohort standardization, Clin-REACT 8B, 14B, and 70B consequently ranked substantially higher in clinical reasoning than in multiple choice and appeared above the equal-standing line, while Clin-REACT 31B remained strong on both benchmark families (Fig. S16B). Together, these findings suggest that medical

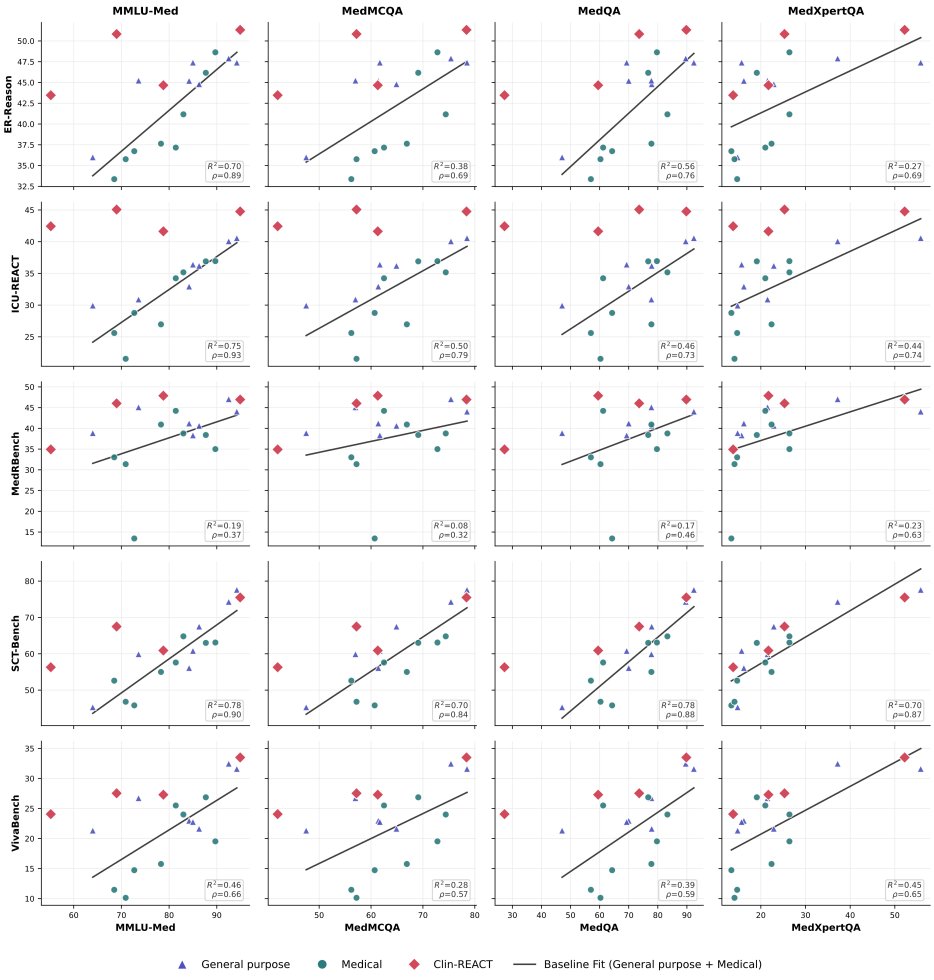


Fig. S17 Benchmark-level relationships between multiple-choice and clinical-reasoning performance. Pairwise scatter plots for all 20 combinations of the five clinical-reasoning benchmarks (rows) and four multiple-choice benchmarks (columns). Each point is one model: general-purpose (triangles), open-source medical (circles), and Clin-REACT (diamonds). Clinical-reasoning benchmark scores are the average of the benchmark's sub-metrics. Within each panel, the solid line is an ordinary least-squares fit to the baseline models only (general-purpose and medical, $n = 15$), with the corresponding coefficient of determination (R^2) and Spearman rank correlation (ρ) inset; Clin-REACT models are overlaid but excluded from the fit and from the reported statistics. Axes are shared within each row and column.

multiple-choice benchmarks capture an important component of medical knowledge but do not consistently represent the broader reasoning capacity required for clinically grounded tasks. Benchmark-level comparisons further supported the observed relationships (Fig. S17).

S10 Data contamination analysis

To assess whether benchmark performance could be influenced by overlap between the Clin-REACT training data and evaluation sets, we conducted complementary semantic-similarity and completion-based contamination analyses (Fig. S18). First, for each sample in ICU-REACT Test and the four external benchmarks (SCT-Bench, ER-Reason, MedRBench, and VivaBench), we identified its nearest neighbor within each of the ICU-REACT Small, Medium, and Large training sets using cosine similarity of the input-output text embeddings. As expected, ICU-REACT Test showed the highest similarity to the ICU-REACT training sets (median cosine similarity, 0.842–0.880), whereas similarity was substantially lower for the external benchmarks: 0.425–0.466 for SCT-Bench, 0.671–0.687 for ER-Reason, 0.493–0.528 for MedRBench, and 0.470–0.486 for VivaBench (Fig. S18A). We subsequently performed a stricter completion-based analysis using Clin-REACT 8B and 70B and their respective Llama 3.1 8B Instruct and Llama 3.3 70B Instruct backbones. For each reference answer, the model was conditioned on prefixes truncated at token positions 10, 15, 20, 25, and 30 of the input, and the subsequent five generated tokens were compared with the corresponding five reference tokens. We defined a sample as exhibiting potential leakage when exact 5-token matches occurred at three or more of the five tested positions [1, 2], thereby requiring repeated continuation-level agreement rather than an isolated phrase match.

Under this criterion, no potential leakage was detected for ICU-REACT, SCT-Bench, ER-Reason, or VivaBench for either Clin-REACT model or either backbone (Fig. S18B). The only non-zero rates occurred on MedRBench and remained low: 0.07% for the Llama 3.1 8B backbone, 0.60% for Clin-REACT 8B, 0.67% for the Llama 3.3 70B backbone, and 0.37% for Clin-REACT 70B. Importantly, the direction of change after fine-tuning was not consistent across scales, with potential leakage increasing relative to the 8B backbone but decreasing relative to the 70B backbone, providing no evidence of a systematic increase attributable to Clin-REACT training. Position-specific exact-match rates were higher than the stricter leakage rates, particularly for MedRBench at later prediction positions, where token-30 match rates ranged from 25.1% to 27.6% across models (Fig. S18C). However, these local matches were also prominent in the pretrained backbones and rarely persisted across enough positions to satisfy the potential-leakage criterion, suggesting that they more likely reflect reproducible or formulaic clinical language than sustained memorization of benchmark answers. Together, the semantic and completion-based analyses provide little evidence that the observed Clin-REACT performance gains were driven by systematic contamination of the evaluated benchmarks.

S11 ICU Variable and Taxonomy Framework

We developed a structured ICU variable and taxonomy framework algined with OMOP concepts to support the building of the ICU-REACT dataset and future model deployment compatible with EHR systems. This framework was designed to reflect clinically meaningful domains consistent with established ICU severity scoring systems and outcome prediction models. Established ICU severity scores

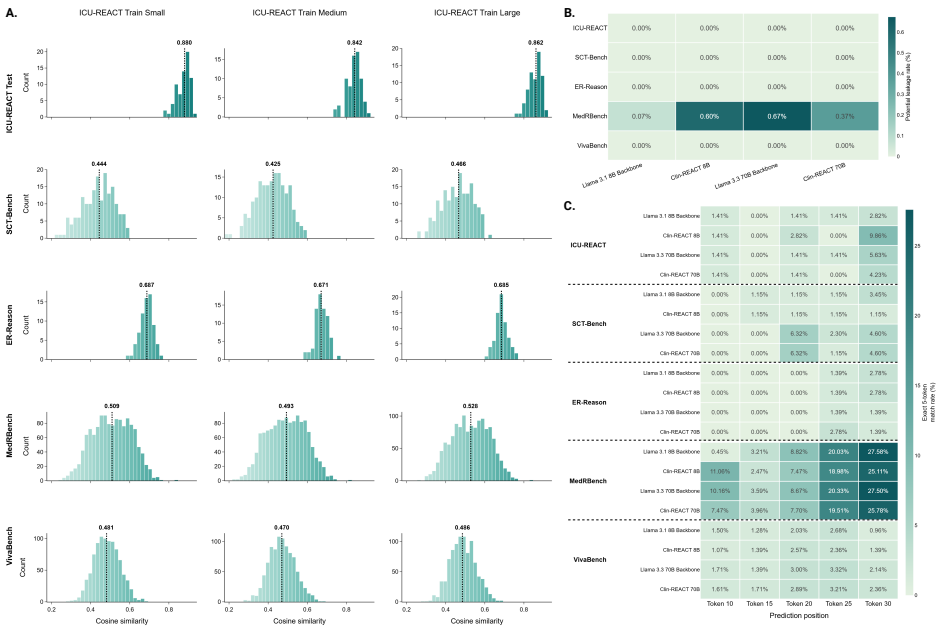


Fig. S18 Data-contamination analysis between ICU-REACT training sets and clinical reasoning evaluation benchmarks. (A) Distributions of nearest-neighbor cosine similarity between samples from five clinical reasoning benchmarks (rows: ICU-REACT Test, SCT-Bench, ER-Reason, MedRBench, and VivaBench) and the ICU-REACT Train Small, Medium, and Large datasets (columns). Bars show the number of evaluation samples within each cosine-similarity interval, with color intensity increasing with similarity; vertical dotted lines indicate the median similarity for each comparison, with the median value reported above each distribution. (B) Potential leakage rates for Clin-REACT 8B and 70B and their corresponding Llama 3.1 8B Instruct and Llama 3.3 70B Instruct backbones across the five clinical reasoning benchmarks. Each cell reports the percentage of samples satisfying the potential-leakage criterion, with darker teal indicating a higher rate. (C) Exact 5-token match rates at five prediction positions (tokens 10, 15, 20, 25, and 30) for the same four models across each benchmark. Rows are grouped by benchmark and separated by dashed horizontal lines; each cell reports the percentage of samples with an exact 5-token match at the corresponding prediction position, with darker teal indicating a higher match rate.

(APACHE, SAPS, SOFA) demonstrate that demographic, diagnostic, physiologic, laboratory, intervention, and admission variables are valuable determinants of patient outcomes [3, 4].

We categorized the organization of variables into eight domains: (1) Diagnosis, (2) Physiology, (3) Laboratory Values, (4) Imaging, (5) Medications, (6) Interventions, (7) Severity Scores and Clinical Assessments, and (8) Administrative Variables. The selection of these domains was guided by medical trainees and informed by prior work defining common data elements and data dictionaries for critical care research [5, 6].

Within each domain, most variables were identified through the common practice of expert elicitation [7], from attending physicians, residents, and medical students with clinical training at the University of Florida College of Medicine. Prior work has defined critical care data elements through expert consensus approaches [5, 7],

while our approach incorporates vignette-based expert elicitation to capture clinician reasoning in context-specific ICU scenarios. Participants were presented with standardized ICU clinical vignettes representing common critical care scenarios. For each vignette, participants selected the variables they considered clinically relevant for patient assessment, monitoring, and management. Responses were collected across participants, and variables endorsed through this process were added to the eight predefined domains.

Variables were consolidated to reduce redundancy and grouped into clinically coherent sub-domains; for instance, the Diagnosis domain was organized into organ system-based groupings (Cardiac, Pulmonary, Neurological, Infectious, Renal, Metabolic, Endocrine, Gastrointestinal, Hepatic, and Hematological) to reflect common etiologies encountered in ICU practice. This expert-driven data collection method aims to ensure that the final framework reflects real-world clinical reasoning and maintains structured organization for future applications. The full list of variables organized by clinical domain and sub-domains can be found on Table S5.

Table S5: ICU variable taxonomy organized by clinical domain and sub-domain.

Domain	**Subcategory**	**Variables**
Diagnosis	Cardiac	acute coronary syndrome; atrial fibrillation; bradycardia; cardiac arrest; cardiogenic shock; complete heart block; endocarditis; heart failure; myocardial infarction; pericardial effusion; supraventricular tachycardia; ventricular fibrillation; ventricular tachycardia; Aortic dissection; Cardiomyopathy; Myocarditis; Valvular heart disease; DVT; Pulmonary Embolism
Diagnosis	Pulmonary	respiratory acidosis; respiratory alkalosis; ventilator-associated pneumonia; ARDS; COPD exacerbation; aspiration pneumonitis; asthma exacerbation; hemothorax; pleural effusion; pneumonia; pneumothorax; pulmonary contusion; pulmonary edema; pulmonary embolism; pulmonary hypertension; respiratory failure; air embolism; fat embolism syndrome; Atelectasis; Pulmonary fibrosis; Bronchiectasis; Lung cancer

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Diagnosis	Neurological	hemorrhage; encephalitis; myxedema coma; Guillain-Barré syndrome; brain death; cerebral edema; encephalopathy; epidural hematoma; intracranial hemorrhage; ischemic stroke; myasthenia gravis; status epilepticus; subarachnoid hemorrhage; subdural hematoma; traumatic brain injury; Spinal cord injury; Brain tumor; Hydrocephalus; ICU-acquired weakness; Delirium (Hypoactive)
Diagnosis	Infectious	bacteremia; catheter-related bloodstream infection; fungemia; necrotizing fasciitis; anaphylaxis; sepsis; septic shock; COVID-19; Influenza; Cellulitis; Abscess; Ventilator-associated pneumonia; CLABSI; CAUTI
Diagnosis	Renal	urinary tract infection; adrenal insufficiency; acute kidney injury; acute tubular necrosis; chronic kidney disease; nephrotic syndrome; rhabdomyolysis; urinary retention; Nephrolithiasis; Glomerulonephritis; Hydronephrosis; Pyelonephritis; Hepatorenal Syndrome
Diagnosis	Metabolic	hyperkalemia; hypernatremia; hypokalemia; hyponatremia; metabolic acidosis; metabolic alkalosis; hyperthermia; hypothermia; malignant hyperthermia; multi-organ dysfunction syndrome; systemic inflammatory response syndrome; Hypomagnesemia; Hypermagnesemia; Hypophosphatemia; Hyperphosphatemia; Heat stroke
Diagnosis	Endocrine	diabetic ketoacidosis; hyperglycemia; hyperosmolar hyperglycemic state; hypoglycemia; thyroid storm; Hypothyroidism; Hyperthyroidism; Cushing's syndrome
Diagnosis	Gastrointestinal	Clostridioides difficile colitis; GI bleed; acute pancreatitis; bowel ischemia; bowel obstruction; cholecystitis; ileus; lower GI bleed; upper GI bleed; meningitis; hemorrhagic stroke; Gastritis; Peptic ulcer disease; Diverticulitis; Volvulus
Diagnosis	Hepatic	acute liver failure; ascites; esophageal varices; liver failure; Cirrhosis; Portal hypertension; Steatohepatitis; Hepatitis

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Diagnosis	Hematological	anemia; deep vein thrombosis; disseminated intravascular coagulation; heparin-induced thrombocytopenia; thrombocytopenia; thrombotic thrombocytopenic purpura; Leukemia; Lymphoma; Sickle cell disease; Hemophilia; Hypovolemic shock
Physiology	Excretion	Urine output; Hourly urine output; Chest tube output; Drain output; Nasogastric tube output; Stool output; Bile output; Jackson-Pratt drain output; Hemovac drain output
Physiology	Fluidics	Fluid intake; Fluid balance; Daily net fluid balance; Cumulative fluid balance; Insensible losses; Bolus volume; Maintenance rate; Cumulative Balance
Physiology	Hemodynamics	Central venous pressure; Cardiac output; Cardiac index; Stroke volume; Stroke volume index; Ejection fraction; Systemic vascular resistance; Systemic vascular resistance index; Pulmonary vascular resistance; Pulmonary vascular resistance index; Pulmonary artery pressure; Pulmonary capillary wedge pressure; Right atrial pressure; Shock index; Stroke volume variation; Pulse pressure variation; Cardiac power output; Fluid responsiveness; Passive leg raise result; IVC distensibility index; Hemodynamic instability; Pulse Pressure
Physiology	Oxygenation	PaO2/FiO2 ratio; Oxygenation index; Alveolar-arterial oxygen gradient; Dead space fraction; Ventilatory ratio; Mixed venous oxygen saturation; Central venous oxygen saturation; Oxygen delivery; Oxygen consumption; Oxygen extraction ratio; End-tidal CO2; Venous-to-arterial CO2 gap; Work of breathing; Spontaneous Breathing Trial (SBT); Rapid shallow breathing index; Diaphragm excursion; pH; PaO2; PaCO2; Bicarbonate; Base excess; Lactate; Anion gap; Mechanical Power; Stress Index; Mean Airway Pressure; Spontaneous Effort; Cough Strength
Physiology	Pressures	Intra-abdominal pressure; Bladder pressure

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Physiology	Ventilation	Mode of ventilation; Respiratory rate (set); Mandatory rate; Spontaneous rate; PEEP; FiO2; Pressure support; Volume control; Inspiratory time; Expiratory time; I:E ratio; Mean airway pressure; Lung volumes; Tidal volume; Minute ventilation; Expiratory minute volume; Peak inspiratory pressure; Plateau pressure; Driving pressure; Auto-PEEP; Mechanical power; Respiratory compliance; Respiratory compliance (static); Respiratory compliance (dynamic); Respiratory system resistance; Respiratory system elastance; Transpulmonary pressure; Airway resistance; Trigger sensitivity
Physiology	Vitals	Mean arterial pressure; Heart rate; Blood pressure systolic; Blood pressure diastolic; Pulse pressure; Respiratory rate; Oxygen saturation; Core temperature; Weight
Imaging	Angiography	Cerebral Angiography; Coronary Angiography; GI bleeding extravasation; Peripheral Angiography
Imaging	Basic Radiography	Abdominal X-ray (KUB); Atelectasis (X-ray); Cardiomegaly; Chest X-ray; Infiltrates; Line position (X-ray); Mediastinal widening; Pelvic X-ray; Pleural effusion (X-ray); Pneumothorax (X-ray); Portable Chest X-ray; Pulmonary edema (X-ray); Tube position (X-ray)
Imaging	CT	Abscess (CT); Aortic dissection (CT); Bowel ischemia (CT); Bowel obstruction (CT); Brain edema; CT Abdomen/Pelvis; CT Chest; CT Head; CT Spine; CTA Chest (PE protocol); CTA Head/Neck; Free air (CT); Intracranial hemorrhage; Midline shift; Pulmonary embolism (CT)
Imaging	EEG	EEG findings
Imaging	EKG	12-lead EKG; EKG Rhythm; Ischemic EKG changes; EKG Findings; ST-segment changes
Imaging	Fluoroscopy	Barium Swallow; Diaphragmatic Fluoroscopy (Sniff Test)
Imaging	MRI	Acute stroke (MRI/DWI); Brain tumor (MRI); Cardiac MRI; MRI Brain; MRI Spine
Imaging	Scintigraphy	Bone Scan; V/Q Scan

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Imaging	Ultrasound	B-lines (Ultrasound); Bedside Echo; Bedside IVC assessment; FAST Exam; Lung Ultrasound; Lung sliding; Abdominal Ultrasound; Bladder Ultrasound; Echocardiogram (TEE); Echocardiogram (TTE); Pericardial effusion findings; Renal Ultrasound; Valvular function findings; Vascular Ultrasound (DVT)
Medications	Analgesics	Acetaminophen; Fentanyl; Hydromorphone; Ketorolac; Methadone; Morphine; Remifentanil
Medications	Antiarrhythmics	Adenosine; Amiodarone; Digoxin; Diltiazem; Esmolol; Lidocaine; Metoprolol
Medications	Antibiotics	Azithromycin; Aztreonam; Cefepime; Ceftriaxone; Daptomycin; Gentamicin; Levofloxacin; Linezolid; Meropenem; Metronidazole; Piperacillin-Tazobactam; Vancomycin
Medications	Anticoagulants	Apixaban; Argatroban; Bivalirudin; Enoxaparin; Heparin; Rivaroxaban; Warfarin
Medications	Antidotes	Andexanet Alfa; Flumazenil; Idarucizumab; Naloxone; Protamine; Sugammadex; Vitamin K
Medications	Antifungals	Amphotericin B; Fluconazole; Micafungin; Voriconazole
Medications	Antihypertensives	Clevidipine; Enalaprilat; Hydralazine; Labetalol; Nicardipine; Nitroglycerin; Nitroprusside
Medications	Antiplatelets	Aspirin; Cangrelor; Clopidogrel; Ticagrelor
Medications	Antivirals	Acyclovir; Ganciclovir; Oseltamivir
Medications	Diuretics	Bumetanide; Furosemide; Metolazone; Spironolactone
Medications	Electrolytes	Calcium Chloride; Calcium Gluconate; Magnesium Sulfate; Potassium Chloride; Sodium Bicarbonate; Sodium Phosphate
Medications	Fluids	Albumin 25%; Albumin 5%; Lactated Ringers; Normal Saline
Medications	Hemostatics	Alteplase; Kcentra; Tranexamic Acid
Medications	Hormones	Desmopressin; Dexamethasone; Glucagon; Hydrocortisone; Insulin; Levothyroxine; Methylprednisolone; Octreotide
Medications	Inotropes	Dobutamine; Dopamine; Isoproterenol; Milrinone
Medications	Paralytics	Cisatracurium; Rocuronium; Succinylcholine; Vecuronium
Medications	Pressors	Angiotensin II; Ephedrine; Epinephrine; Norepinephrine; Phenylephrine; Vasopressin

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Medications	Protectants	Famotidine; Metoclopramide; Ondansetron; Pantoprazole
Medications	Reversal/Anecdotes	Sugammadex; Vitamin K
Medications	Sedatives	Dexmedetomidine; Etomidate; Ketamine; Lorazepam; Midazolam; Propofol
Medications	Supportive	Alteplase; Prothrombin Complex Concentrate; Tranexamic Acid
Administrative	Ethics	Advance directives; DNR/DNI status; Family meeting notes; Hospice referral
Administrative	Habits	Alcohol use; Drug use history; Smoking status
Administrative	Identity	Age; Date of birth; Ethnicity; Gender; Height; Language; Medical record number; Name; Race
Administrative	Logistics	Admission date; Admission diagnosis; Admission source; Admission time; Admitting service; Discharge date; Hospital length of stay; ICU admission date; ICU length of stay
Administrative	Notes	Nutrition notes; Occupational therapy notes; Palliative care notes; Physical therapy notes; Respiratory therapy notes; Social work notes; Speech therapy notes
Administrative	Social	Employment status; Housing status; Insurance type; Next of kin
Administrative	Status	Allergies; BMI; Organ donor status; Pregnancy status; Vaccination status
Intervention	Advanced Life Support	ECMO; IABP; VAD; continuous renal replacement therapy; dialysis; hemodialysis; implantable cardioverter-defibrillator; peritoneal dialysis; permanent pacemaker; temporary pacemaker; therapeutic hypothermia
Intervention	Blood Product Transfusion	cryoprecipitate transfusion; plasma transfusion; platelet transfusion; red blood cell transfusion
Intervention	Drainage & Tubes	Foley catheter; Hemovac drain; Jackson-Pratt drain; biliary drain; chest tube placement; nasogastric tube; orogastric tube; percutaneous endoscopic gastrostomy; rectal tube; suprapubic catheter; wound VAC
Intervention	Invasive Procedures	cardiac catheterization; cardioversion; colonoscopy; debridement; defibrillation; endoscopy; lumbar puncture; paracentesis; percutaneous coronary intervention; pericardiocentesis; skin graft; thoracentesis

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Intervention	Neurological Interventions	external ventricular drain; intracranial pressure monitor; lumbar drain; train-of-four monitoring
Intervention	Nutrition & Fluid Management	IV fluids administered; enteral feeding rate; enteral nutrition; fluid resuscitation; parenteral nutrition
Intervention	Procedures	tracheostomy care
Intervention	Prophylaxis & Supportive Care	DVT; active dialysis orders; bladder scan; current oxygen delivery device; dvt prophylaxis; neuromuscular blockade; neuromuscular electrical stimulation; pressure ulcer staging; sedation vacation; sequential compression device; spontaneous awakening trial; stress ulcer prophylaxis; tracheostomy care; venous thromboembolism prophylaxis
Intervention	Respiratory Interventions	bilevel positive airway pressure; bronchoscopy; continuous positive airway pressure; extubation; high-flow nasal cannula; intubation; mechanical insufflation-exsufflation; oxygen via tracheostomy collar; prone positioning; recruitment maneuvers; reintubation; tracheostomy
Intervention	Vascular Access	PICC line; arterial line; central line; epidural catheter; intraosseous line; midline catheter; peripheral IV; pulmonary artery catheter
Laboratory Measurements	Cardiac	BNP; CK; CK-MB; NT-proBNP; Troponin I; Troponin T
Laboratory Measurements	Coagulation	Activated clotting time (ACT); Anti-Xa level; Antithrombin III; D-dimer; Fibrinogen; INR; PT; PTT; Protein C activity; Protein S activity; ROTEM; TEG - MA; TEG - R time; Thrombin time
Laboratory Measurements	Complete Blood Count	Basophils (absolute); Eosinophils (absolute); Hematocrit; Hemoglobin; Immature granulocytes (Bands); Lymphocytes (absolute); MCH; MCHC; MCV; Monocytes (absolute); Neutrophils (absolute); Nucleated RBCs; Peripheral smear; Platelet count; RBC count; RDW; Reticulocyte count; WBC count
Laboratory Measurements	Comprehensive Metabolic Panel	ALP; ALT; AST; Albumin; Anion gap; BUN; Bicarbonate; Calcium (total); Chloride; Creatinine; Direct Bilirubin; Glucose; Indirect Bilirubin; Ionized Calcium; Magnesium; Phosphorus; Potassium; Sodium; Total Bilirubin; Total Protein; Uric acid

Continued on next page

Table S5 – continued from previous page

Domain	Subcategory	Variables
Laboratory Measurements	Endocrine	ACTH; Cortisol (random); Free T3; Free T4; HbA1c; Insulin level; PTH; TSH
Laboratory Measurements	Inflammatory	CRP; ESR; Ferritin; Haptoglobin; Interleukin-6 (IL-6); LDH
Laboratory Measurements	Microbiology	Blood culture; C. diff PCR; CSF culture; Fungal culture; Gram stain; MRSA Screen; Procalcitonin; Sputum culture; Urine culture; Viral PCR (Respiratory Panel); Wound culture
Laboratory Measurements	Nutritional	Cystatin C; Folate; Prealbumin; Transferrin; Vitamin B12; Vitamin D
Laboratory Measurements	Toxicology	Acetaminophen level; Amikacin level; Carbamazepine level; Cyclosporine level; Digoxin level; Ethanol level; Gentamicin level; Lithium level; Phenytoin level; Salicylate level; Tacrolimus level; Tobramycin level; Urine drug screen; Valproic acid level; Vancomycin trough
Laboratory Measurements	Urinalysis	Urine bilirubin; Urine blood; Urine glucose; Urine ketones; Urine leukocyte esterase; Urine microscopic exam; Urine nitrite; Urine osmolality; Urine pH; Urine protein; Urine specific gravity
Scores and Assessments	Acute severity of illness scores	APACHE II score; LODS; OASIS; SAPS II; SOFA score; qSOFA
Scores and Assessments	Chronic & Comorbidity Indices	Charlson Comorbidity Index; Clinical Frailty Scale; Elixhauser Comorbidity Index
Scores and Assessments	Neuro-Cognitive Assessments	FOUR Score; Glasgow Coma Scale; Hunt and Hess Scale; Mental status; Muscle strength; NIH Stroke Scale; Pupil response; Seizures; modified Rankin Scale
Scores and Assessments	Nutrition Scores	MUST; NUTRIC score
Scores and Assessments	Organ Function Scores	Balthazar score; Child-Pugh score; MELD score
Scores and Assessments	Pain Scores	BPS; CPOT; NVPS; Pain score
Scores and Assessments	Physical Exam Findings	Abdominal distension; Bowel sounds; Capillary refill; Peripheral edema; Skin turgor
Scores and Assessments	Psychiatric Assessments	CAM-ICU; Delirium
Scores and Assessments	Safety Risk Assessments	Braden Scale; CURB-65; MEWS; Morse Fall Scale; NEWS; Wells score
Scores and Assessments	Sedation Scales	RASS; SAS; Sedation level

S12 ICU Topics Dimensions

Prompting a language model repeatedly with only a topic label (e.g. "septic shock") yields cases that vary in wording but converge on a small number of prototypical presentations, which limits the diversity gained from augmentation. To avoid this, we decomposed each ICU topic into four axes along which real cases differ: the underlying clinical scenario (the etiology or precipitating event), the presentation (the bedside findings or trigger prompting a decision), the management decision under consideration, and contextual modifiers (comorbidities, competing risks, resource constraints, and goals-of-care limitations that alter what is appropriate). These axes are largely independent of one another: the same presentation can arise from different etiologies, and the same decision can be appropriate or contraindicated depending on the modifier. Sampling them separately therefore produces variation that is clinically meaningful rather than merely lexical, and lets a modest number of enumerated items cover a much larger space of distinct cases.

We selected nine clinician-informed topics spanning the decision contexts that dominate adult ICU care, and for each topic used ChatGPT to enumerate candidate items along the four axes, seeded with the topic definition and constrained to decisions that arise in routine intensive care. The generated lists were then reviewed and edited to remove redundant or implausible entries and to keep items at a comparable level of granularity. The final set comprises 315 items: 96 clinical scenarios, 84 presentations, 87 management decisions, and 48 contextual modifiers, distributed across topics as shown in Table S6. Drawing a single item from each axis would yield 50,017 distinct scenario specifications across the nine topics, which allowed for diverse generation of training samples.

Table S6: ICU clinical topics and dimensions used to generate ICU-REACT training sets

Topic	Dimension	Items
Hemodynamic Instability / Shock	Clinical scenarios	Septic shock from pneumonia; Septic shock from intra-abdominal source; Postoperative hypotension after major surgery; Acute upper GI hemorrhage with volume loss; Traumatic hemorrhagic shock; Cardiogenic shock after acute MI; Acute decompensated heart failure with low output; Acute right-ventricular failure from massive PE; Cardiac tamponade with obstructive physiology; Anaphylactic distributive shock; Neurogenic/spinal shock after trauma; Adrenal/ relative cortisol insufficiency in shock; Tension pneumothorax causing obstructive shock; Dialysis/CRRT-associated hypotension in ICU

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Hemodynamic Instability / Shock	Presentations	Sustained low MAP despite initial fluids; Rising lactate and signs of poor perfusion; Cool, mottled extremities and delayed cap refill; Warm, vasodilatory shock with wide pulse pressure; Oliguria or declining urine output; Acute change in mental status from hypoperfusion; Tachycardia with narrow pulse pressure; Escalating vasopressor requirement over hours; Hypotension with new arrhythmia; Need for closer hemodynamic monitoring; Hypotension in setting of limited fluid tolerance (ARDS/CHF)
Hemodynamic Instability / Shock	Management decisions	Initiate norepinephrine as first-line vasopressor; Add a second vasopressor such as vasopressin; Start inotropic support for low cardiac output; Administer another fluid bolus vs hold further fluids; Transfuse packed RBCs for hemorrhagic component; Obtain bedside echocardiography to define etiology; Place arterial and/or central line for monitoring; Initiate stress-dose corticosteroids for refractory shock; Arrange source-control procedure for shock driver; Escalate to mechanical/advanced circulatory support; Begin conservative fluid strategy after stabilization
Hemodynamic Instability / Shock	Contextual modifiers	Patient has significant cardiopulmonary comorbidities; High risk of fluid overload/ARDS so resuscitate cautiously; Limited vascular access or line-related infection concern; Post-cardiac surgery physiology alters pressor choice; Immunocompromised host with atypical infection; Goals-of-care or DNR/DNI status limits escalation

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Respiratory Failure	Clinical scenarios	ARDS secondary to severe pneumonia; Community-acquired pneumonia with hypoxemic failure; COPD exacerbation with hypercapnic respiratory failure; Obesity hypoventilation/OSA causing ventilatory failure; Postoperative atelectasis with impaired gas exchange; Aspiration event in intubated patient; Ventilator-associated pneumonia in ICU; Neuromuscular weakness (GBS/MG) causing hypoventilation; Cardiogenic pulmonary edema with respiratory compromise; Prolonged mechanical ventilation needing weaning; Post-extubation stridor or upper-airway compromise; Trauma/chest wall injury with poor ventilation; Severe asthma in ICU not responding to usual therapy
Respiratory Failure	Presentations	Worsening oxygen requirement despite current support; Increased work of breathing and tachypnea; Rising CO2 with somnolence or confusion; Ventilator dyssynchrony or high peak pressures; Difficulty clearing secretions or poor cough; Failure of HFNC/NIV trial; Hemodynamic instability related to high PEEP; Ready for spontaneous breathing trial; Failed prior extubation attempt; Recurrent desaturation with minimal activity; Concern for aspiration or VAP on current ventilator settings
Respiratory Failure	Management decisions	Escalate to non-invasive ventilation (e.g. BiPAP); Proceed with endotracheal intubation and mechanical ventilation; Adjust ventilator settings for lung-protective strategy; Increase or titrate PEEP/FiO2 to maintain oxygenation; Initiate prone positioning or recruitment maneuvers; Start weaning protocol/SBT and assess extubation readiness; Switch to HFNC from NIV for better tolerance; Request bronchoscopy/ suctioning for secretion management; Consult for tracheostomy due to prolonged ventilation; Add sedation/analgesia to improve ventilator synchrony; Treat underlying cardiac/pulmonary edema to improve oxygenation

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Respiratory Failure	Contextual modifiers	Patient is hemodynamically fragile and cannot tolerate high PEEP; High aspiration risk/enteral feeding ongoing; Difficult airway or prior failed intubation; Immunosuppressed patient at risk for atypical infections; Obesity/body habitus complicates ventilation strategy; Infection control/isolation requirements limit procedures
Sepsis and Severe Infections	Clinical scenarios	Pneumonia with sepsis requiring ICU admission; Complicated intra-abdominal infection/ postoperative peritonitis; Central line-associated bloodstream infection; Urinary tract infection progressing to urosepsis; Skin/soft-tissue infection or necrotizing fasciitis; Infected indwelling device or prosthesis; Febrile neutropenia with suspected bacterial source; Ventilator-associated pneumonia in intubated patient; Biliary tract infection/cholangitis; Catheter-associated peritonitis in dialysis patient; Endocarditis suspected in ICU patient; Sepsis of unclear source requiring broad coverage
Sepsis and Severe Infections	Presentations	Fever or hypothermia with leukocytosis; Hypotension or rising vasopressor needs; New or worsening organ dysfunction (renal, hepatic, neuro); Persistent or recurrent bacteremia on cultures; Local signs of infection at catheter or wound site; Poor response to initial antibiotics; Elevated inflammatory markers or rising lactate; Respiratory decompensation due to infection; Need for source control identified on imaging; Immunocompromised host with atypical presentation
Sepsis and Severe Infections	Management decisions	Initiate broad-spectrum empiric antibiotics; De-escalate antibiotics based on culture and sensitivity; Remove or exchange an infected central line or device; Arrange surgical or percutaneous source control; Start or escalate vasopressors for septic shock; Add stress-dose steroids for refractory septic shock; Extend or shorten duration of antibiotic therapy; Switch to antifungal or antiviral coverage when indicated; Isolate patient and apply infection control measures; Initiate early enteral nutrition in septic patient; Consult infectious diseases for complex resistant organisms

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Sepsis and Severe Infections	Contextual modifiers	Renal or hepatic dysfunction limiting antibiotic choice; Recent hospitalization or MDR organism risk; Pregnancy/postpartum infection considerations; Neutropenia/oncology patient with low inflammatory response; Source not yet identified; diagnostics still pending; Coagulopathy/bleeding risk complicating source-control procedure
Neurological Emergencies	Clinical scenarios	Traumatic brain injury after fall or MVC; Spontaneous intracranial hemorrhage; Acute ischemic stroke requiring ICU monitoring; Status epilepticus or recurrent seizures; Post-ictal coma not awakening as expected; ICU delirium in mechanically ventilated patient; Sedation-related depressed mental status; Suspected increased ICP from mass effect or edema; Neurosurgical postoperative patient with neuro changes; Meningitis/encephalitis with altered mental status
Neurological Emergencies	Presentations	Sudden drop in GCS or unresponsiveness; Unequal or sluggish pupils; Ongoing or recurrent seizure activity; New agitation or hyperactive delirium; Failure to awaken after sedation wean; Headache and vomiting suggesting increased ICP; Hypertension/bradycardia pattern concerning for herniation; Speech or focal motor deficits; Fever with meningismus in altered patient
Neurological Emergencies	Management decisions	Secure airway/intubate to protect from aspiration; Obtain emergent neuroimaging (CT/MRI/CTA); Initiate ICP-lowering measures (head-up, mannitol, HTS); Start or escalate antiepileptic medication; Adjust or lighten sedation to assess neuro status; Initiate delirium screening and non-pharm management; Control blood pressure within neuro-targeted range; Consult neurosurgery/neurology urgently; Place invasive monitoring (ICP/EVD) when indicated
Neurological Emergencies	Contextual modifiers	Anticoagulated/antiplatelet patient with bleeding risk; Cervical spine precautions or trauma constraints; Concurrent sepsis/hemodynamic instability; Renal/hepatic impairment limiting sedative choice; Difficult family/goals-of-care situation around prognosis

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Renal Failure / Electrolyte Disorders	Clinical scenarios	Sepsis-associated AKI in ICU; Postoperative AKI after major surgery; Contrast-induced nephropathy; Cardiorenal syndrome with volume overload; Acute tubular necrosis from hypotension; Drug-induced nephrotoxicity; Obstructive uropathy recognized in ICU patient; Tumor lysis/metabolic ICU patient; Isolated severe hyperkalemia; Hypo/hypernatremia related to ICU therapies; CRRT-dependent patient with circuit or prescription issues
Renal Failure / Electrolyte Disorders	Presentations	Oliguria or anuria over several hours; Rising creatinine/BUN compared to baseline; Pulmonary/peripheral edema from fluid overload; Metabolic acidosis not improving; Dangerous electrolyte derangement (qualitative); Hemodynamic instability limiting diuresis; Need to start nephrotoxic/renally-cleared drug; CRRT filter clotting or inadequate clearance; Uremic symptoms or concerns in ICU patient
Renal Failure / Electrolyte Disorders	Management decisions	Initiate CRRT or intermittent hemodialysis; Temporize and treat hyperkalemia/electrolyte abnormality; Restrict fluids and use diuretics for volume management; Adjust or hold nephrotoxic/renally-cleared medications; Change CRRT prescription (dose, fluid removal, anticoagulation); Investigate and relieve obstruction; Consult nephrology for AKI not improving; Optimize hemodynamics to improve renal perfusion; Plan earlier RRT due to multi-organ failure; Correct sodium or acid-base disturbance cautiously
Renal Failure / Electrolyte Disorders	Contextual modifiers	Hemodynamic fragility limiting fluid removal; Liver disease or coagulopathy affecting dialysis access; Pediatric/small adult size affecting modality choice; Active sepsis or catheter infection risk; Goals of care limit initiation of chronic dialysis

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Cardiac Emergencies	Clinical scenarios	New-onset atrial fibrillation with rapid ventricular response; Sustained ventricular tachycardia or VF arrest in ICU; Acute coronary syndrome in hemodynamically unstable patient; Post-cardiac arrest (ROSC) patient in ICU; Symptomatic bradycardia or high-grade AV block; Pericardial tamponade physiology; Post–cardiac surgery patient with arrhythmia; Decompensated heart failure with pulmonary edema; Takotsubo/stress cardiomyopathy presentation; Electrolyte-triggered arrhythmia in ICU
Cardiac Emergencies	Presentations	Arrhythmia with hypotension or chest discomfort; Recurrent or shockable rhythm post-ROSC; Persistent ST changes or ischemic symptoms; Low cardiac output with cool extremities; Syncope or presyncope in monitored patient; Rising filling pressures/JVD in ICU; Worsening dyspnea/pulmonary edema on monitor; Need for rate control before procedure; Bradycardia not responding to atropine
Cardiac Emergencies	Management decisions	Perform synchronized cardioversion or defibrillation; Choose rate vs rhythm control for AF; Start or adjust anticoagulation for arrhythmia; Optimize hemodynamics post-ROSC (fluids/pressors); Place temporary pacemaker or consult EP; Treat underlying ischemia/ACS and call cardiology; Diurese and afterload reduce in decompensated HF; Pericardiocentesis/ pericardial drain for tamponade; Correct electrolytes/precipitating factors
Cardiac Emergencies	Contextual modifiers	Recent surgery/bleeding risk affecting anticoagulation; Chronic arrhythmia with different baseline HR goals; Renal dysfunction affecting medication choices; Do-not-shock or limited-resuscitation status; Limited catheterization/surgical resources overnight

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Sedation, Pain, and Delirium Management	Clinical scenarios	Mechanically ventilated patient on continuous sedation; Trauma/TBI patient with agitation; Alcohol or benzodiazepine withdrawal in ICU; Postoperative patient with significant pain needs; Septic patient developing ICU delirium; Long-stay ICU patient with sleep disruption; Patient failing ventilator weaning due to oversedation; Elderly patient with hypoactive delirium; Patient with history of psychiatric disease on home meds
Sedation, Pain, and Delirium Management	Presentations	Agitation and RASS above target; Inadequate analgesia despite current regimen; Fluctuating confusion or CAM-ICU positive; Failure to awaken after sedation interruption; Autonomic hyperactivity suggesting withdrawal; Nighttime agitation/sundowning pattern; Ventilator dyssynchrony due to under/over-sedation; Family concern about over-sedation; Need for neuro exam but sedation too deep
Sedation, Pain, and Delirium Management	Management decisions	Adjust sedative/analgesic doses or agents; Initiate or repeat delirium screening and non-pharm measures; Treat alcohol/benzo withdrawal with appropriate protocol; Lighten sedation to facilitate weaning/mobilization; Add antipsychotic or alpha-2 agonist for agitation; Schedule pain medications instead of PRN only; Institute daily sedation interruption; Consult psych/pain service for complex regimen; Modify ventilator or sedation to improve synchrony
Sedation, Pain, and Delirium Management	Contextual modifiers	Renal/hepatic impairment limiting drug choice; Older adult/fall risk with sedative sensitivity; History of substance use disorder/tolerance; Concomitant sepsis/organ failure changing metabolism; Goals-of-care emphasizing comfort over wakefulness

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Nutrition and Metabolic Support	Clinical scenarios	Intubated patient NPO requiring early enteral nutrition; Postoperative abdominal surgery with poor gastric tolerance; Severe pancreatitis with high aspiration risk; Septic/catabolic patient needing high protein/calorie support; ICU patient with uncontrolled hyperglycemia; Patient unable to maintain enteral access (dislodged tube); Prolonged ICU stay with inadequate oral intake; Burn/trauma patient with elevated metabolic demands; Patient with refeeding risk/ malnutrition on admission
Nutrition and Metabolic Support	Presentations	High gastric residuals or vomiting with enteral feeds; Prolonged NPO status due to procedures; Concern for aspiration on current feeding route; Wide glucose variability or persistent hyperglycemia; Electrolyte shifts after starting nutrition; Inability to meet calorie goals with current regimen; Need to minimize fluid volume in feeds; Feeding interruptions due to imaging/ procedures
Nutrition and Metabolic Support	Management decisions	Start or advance enteral nutrition; Switch to post-pyloric or jejunal feeding route; Initiate parenteral nutrition due to intolerance of enteral feeds; Adjust insulin regimen or initiate infusion for glycemic control; Modify caloric/protein goals based on catabolic state; Implement refeeding monitoring and electrolyte replacement; Use trophic/low-volume feeds when full feeds not tolerated; Coordinate feeding schedule around procedures; Consult nutrition/metabolic support team
Nutrition and Metabolic Support	Contextual modifiers	Renal/hepatic dysfunction affecting formula choice; Need for strict fluid restriction; Obesity/ underweight requiring adjusted calculations; Immunocompromised or oncology patient; High aspiration risk/ventilated patient positioning
Hematologic / Coagulation Issues	Clinical scenarios	Active upper or lower GI bleeding; Postoperative surgical site bleeding; Coagulopathy from liver disease or sepsis; ICU patient on warfarin/DOAC with bleeding; Thrombocytopenia or suspected HIT; ICU patient at high risk for VTE due to immobility; DIC in setting of sepsis/trauma; Planned invasive procedure requiring correction

Continued on next page

Table S6 – continued from previous page

Topic	Dimension	Items
Hematologic / Coagulation Issues	Presentations	Ongoing oozing from line/wound sites; Drop in hemoglobin/hematocrit over short interval; Abnormal coagulation studies needing correction; Bleeding while on therapeutic anticoagulation; Need to start VTE prophylaxis but bleeding risk present; New thrombotic event while platelets are low; Multiorgan failure with DIC picture; Upcoming procedure with high bleeding risk
Hematologic / Coagulation Issues	Management decisions	Reverse or hold anticoagulation/antiplatelet therapy; Transfuse PRBCs/platelets/plasma/cryo as indicated; Start or withhold pharmacologic VTE prophylaxis; Evaluate and treat suspected HIT; Correct coagulopathy before procedures; Balance bleeding vs clotting in DIC/sepsis; Consult hematology for complex coagulopathy; Restart anticoagulation when bleeding controlled
Hematologic / Coagulation Issues	Contextual modifiers	Renal/hepatic impairment affecting drug clearance; Recent surgery with strict surgeon preferences; Limited blood product availability; High thrombotic risk (mechanical valve, recent PE); Palliative/comfort-focused goals of care

S13 Prompts

S13.1 Seed generation prompts

Prompt 1

Prompt for initial generation of ICU-REACT decision-making questions and retrieval tasks.

You are a medical expert. You have access to an electronic health record (EHR) database.
Generate `{n_questions}` decision-making example questions that can be answered using an EHR database.

Each question should contain each of the following elements:

1. **Question that can be answered using an EHR database.** The question should meet the following requirements:
   - Should be in the context of an Intensive Care Unit (ICU) setting.
   - Should be actionable and discrete, focusing on decision-making that physicians would consider.
   - Should reference one or more tables that would be found in an EHR database.
   - Should be limited to individual patients.
   - Should be applicable to most ICU unit types.
   - Should cover different organ systems and topics relevant to the ICU setting.
   - Can have multiple retrieval tasks from multiple tables.
   - Should be unique and not highly similar to the examples provided.
   - Should be formulated without restating the context of the question, but rather as a direct inquiry.
   - Should be concise and focused, avoiding unnecessary complexity or jargon.
   - Should be a single question, not a compound question or multiple questions in one.
   - Should not ask about value trends or changes over time, as the focus is on discrete, actionable inquiries.
   - Should not ask about current status of a patient, but rather about specific actions or decisions that can be made based on the available data.
2. **Context in which the question is asked.** The context should follow the template below:

   ```
   [age][gender] patient with [past_medical_history] admitted to the ICU due to
   [surgery_or_current_complication] is developing [critical_condition]
   ```

3. **Category of the question.** This should be a list of tags for the question, such as cardiovascular, respiratory, renal, or other relevant ICU categories.
4. **ICU topic.** This should specify the relevant ICU topic that the question belongs to.
5. **Retrieval tasks that need to be performed to answer the question.** Each task should include:
   - Category of data that needs to be retrieved.
   - List of variables to retrieve for the corresponding task. Do not include "dose", "route", "frequency", "timestamp", "start time", or similar fields in the variables list.
6. **Reasoning behind the retrieval tasks needed to answer the question.** This should be a step-by-step explanation of why each variable is important in the context of the question.

Examples of retrieval tasks:

- **Task:** Patient demographics
  Variables: [age, gender, race, ethnicity]
- **Task:** Hematology test results
  Variables: [WBC count, RBC count, Hemoglobin, Hematocrit, Platelet count]
- **Task:** Arterial blood gas (ABG) results
  Variables: [pH, pCO2, pO2, HCO3, SaO2, lactate]
- **Task:** Vital signs
  Variables: [heart rate, respiratory rate, blood pressure systolic, blood pressure diastolic, temperature, oxygen saturation]

- **Task:** Ventilator settings
  Variables: [FiO2, PEEP, tidal volume, respiratory rate set, plateau pressure, mode of ventilation]
- **Task:** Renal function labs
  Variables: [creatinine, BUN, urine output, eGFR]
- **Task:** Sepsis-related labs
  Variables: [procalcitonin, CRP, lactate, WBC count, temperature, blood cultures]
- **Task:** Vasopressor administration
  Variables: [norepinephrine, epinephrine, vasopressin, dobutamine]
- **Task:** Sedation and analgesia medications
  Variables: [propofol, midazolam, fentanyl, dexmedetomidine]
- **Task:** ICU admission details
  Variables: [ICU admission time, admitting service, ICU type, primary diagnosis]

Output format:
Generate a JSON object with the generated examples following this template:

```
{
  "examples": [
    {
      "question": "User question",
      "context": "Patient context for the question",
      "category": "List of tags for the question",
      "icu_topic": "Relevant ICU topic the question belongs to",
      "retrieval_tasks": [
        {
          "task": "Task to retrieve information from the database",
          "variables": "List of variables to retrieve for corresponding task"
        }
      ],
      "reasoning": "Reasoning behind the question and corresponding retrieval tasks."
    }
  ]
}
```

Few-shot examples:
Here are some examples of questions in JSON format:

```
{examples}
```

Prompt 2

Prompt for mapping ICU-REACT clinical questions to high-level ICU topics.

You are an ICU attending physician. Your task is to categorize clinical questions and their patient context, if provided, into high-level ICU topics.

ICU topic list. Use these labels exactly:

1. **Hemodynamic Instability / Shock**
   Types: Septic, hypovolemic, cardiogenic, obstructive, such as pulmonary embolism or tamponade.
   Management:
   - Rapid fluid resuscitation and monitoring of response.
   - Use of vasopressors, with norepinephrine being the most common.
   - Inotropes for impaired cardiac output.
   - Monitoring lactate, central venous pressure, and echocardiography.
2. **Respiratory Failure**
   Types: Hypoxemic, such as acute respiratory distress syndrome or pneumonia; and hypercapnic, such as chronic obstructive pulmonary disease or neuromuscular weakness.
   Management:

- Oxygen supplementation, high-flow nasal cannula, or non-invasive ventilation.
- Intubation and mechanical ventilation.
- Adjusting ventilator settings, including FiO_2, PEEP, and tidal volume.
- Weaning and extubation planning.

3. **Sepsis and Severe Infections**
   Management:
   - Early recognition using qSOFA or SOFA.
   - Blood cultures and broad-spectrum antibiotics.
   - Source control, such as line removal or abscess drainage.
   - Hemodynamic support with fluids and vasopressors.
4. **Neurological Emergencies**
   Scenarios: Stroke, intracranial hemorrhage, traumatic brain injury, seizures, delirium, and coma.
   Management:
   - Neurological checks, including GCS, RASS, and pupillary response.
   - Intracranial pressure monitoring and osmotic therapy, such as mannitol or hypertonic saline.
   - Airway protection in reduced consciousness.
   - Sedation and delirium management.
5. **Renal Failure / Electrolyte Disorders**
   Management:
   - Monitor urine output and creatinine.
   - Correct electrolyte imbalances, such as hyperkalemia and hyponatremia.
   - Initiate renal replacement therapy, including CRRT or dialysis.
   - Optimize fluid balance.
6. **Cardiac Emergencies**
   Scenarios: Arrhythmias, myocardial infarction, and post-cardiac arrest care.
   Management:
   - ACLS protocols for arrest.
   - Antiarrhythmics, such as amiodarone or lidocaine.
   - Defibrillation or pacing.
   - Hemodynamic optimization after return of spontaneous circulation.
7. **Sedation, Pain, and Delirium Management**
   Management:
   - Balance sedation to ensure comfort while preserving wakefulness.
   - Sedatives, including propofol, dexmedetomidine, and midazolam, and analgesics, including opioids.
   - Delirium screening using CAM-ICU.
   - Early mobilization and sleep hygiene.
8. **Nutrition and Metabolic Support**
   Management:
   - Enteral nutrition, such as nasogastric or post-pyloric feeding.
   - Parenteral nutrition if enteral nutrition is not feasible.
   - Glycemic control while avoiding hypoglycemia and hyperglycemia.
9. **Hematologic / Coagulation Issues**
   Scenarios: Bleeding, thrombosis, and anticoagulation management.
   Management:
   - Transfusion support, including packed red blood cells, platelets, and plasma.
   - Deep vein thrombosis prophylaxis, such as heparin or compression devices.
   - Reversal of anticoagulants in bleeding.

Rules:

- Always choose exactly one primary topic and return it under the key `topic`.
- Optionally choose any number of additional, clearly relevant topics and return them under the key `secondary_topics`.
- Both `topic` and every entry in `secondary_topics` must be copied from the topic titles above verbatim. Case and punctuation may differ, but the text should match.
- If the question is very general but clearly fits one domain, pick the best single topic.

- If multiple domains are clearly involved, pick the most central one as `topic` and add the others to `secondary_topics`.

Return strict JSON only, with this schema:

```
{
  "topic": "<one of the 9 topic titles>",
  "secondary_topics": ["<other topic title>", "..."]
}
```

Patient context, which may be empty:

```
{context}
```

Clinical question:

```
{question}
```

Prompt 3

Prompt for classifying clinician feedback on ICU-REACT samples.

You are an expert ICU clinician-educator and annotation quality assurance analyst.
You will be shown one ICU-REACT sample and one clinician feedback comment about that sample.
Your task is to classify the clinician feedback into exactly one of the following categories:

1. `question_context_improvement`
   Use this when the feedback is mainly about improving the patient context, the framing of the question, missing clinical state details, time anchoring, trajectory, care supports, realism, consistency, or making the decision fork clearer.
2. `variable_selection`
   Use this when the feedback is mainly about which variables should or should not be retrieved, missing key discriminating variables, irrelevant variables, over-selection, under-selection, or the need for different data elements.
3. `reasoning_improvement`
   Use this when the feedback is mainly about how the rationale or reasoning should be written, what logic should be explained, how variables should be connected to the decision, missing interpretation, weak justification, or reasoning structure.
4. `unclassified`
   Use this when the feedback is not clearly about any of the above categories, is too vague, or does not contain actionable critique.

Return only valid JSON with exactly these keys:

```
{
  "label": "question_context_improvement" | "variable_selection" | "reasoning_improvement" |
     "unclassified",
  "confidence": 0.0,
  "rationale": "brief explanation"
}
```

Rules:

- Choose exactly one label.
- Do not output markdown.
- Confidence must be between 0 and 1.
- Base your decision primarily on the clinician feedback, using the sample only as context.

Classify the following clinician feedback.

```
sample_id: {sample_id}
reviewer_label: {reviewer_label}
question: {question}
context: {context}
selected_variables: {selected_variables}
model_reasoning_text: {model_reasoning_text}
clinician_feedback:
{clinician_feedback}

Return only the JSON object.
```

Prompt 4

Prompt for refining reasoning on ICU-REACT train seed samples using clinician feedback.

You are an expert ICU clinician-educator improving ICU-REACT reasoning samples.
You will be given:

- the current clinical question,
- the current patient context,
- the current reasoning text,
- clinician feedback comments that were already classified as `REASONING IMPROVEMENT` feedback.

Your task is to improve only the reasoning.

Core objective:
Turn incomplete, unfocused, weakly justified, or clinically imprecise reasoning into stronger clinician-style reasoning that better uses the available context, question, and clinician feedback.

Important principles:

1. Preserve the original clinical intent unless the clinician feedback clearly indicates the reasoning is framed incorrectly.
2. Do not modify the question.
3. Do not modify the context.
4. Improve the reasoning so it better explains:
   - why the available clinical information matters for the decision or interpretation,
   - how it relates to the specific clinical question,
   - what trajectories, contraindications, alternative explanations, or missing discriminators should matter,
   - and how a chart-reviewing clinician would think through the decision.
5. The reasoning should read like one coherent clinician-style explanatory paragraph, not bullets or JSON-like fragments.
6. Do not invent unsupported facts, lab values, vitals, events, treatments, imaging findings, or outcomes that are not present in the sample or directly justified by the context or question.
7. Do not simply restate the context; explain its clinical relevance.
8. If the clinician feedback clearly implies more than one clinically distinct interpretation of how the reasoning should be strengthened, you may create multiple refinement paths.
9. Prefer a small number of high-quality refinement paths over many weak ones.

When to create multiple refinement paths:

- Create multiple paths only if the clinician feedback clearly supports distinct reasoning framings that would materially change the explanation.
- For example, if the feedback indicates the reasoning should differ depending on whether the main concern is severity assessment, treatment response, contraindication screening, or distinguishing competing etiologies, separate these into distinct reasoning paths only if they truly represent different coherent reasoning directions.

- If the feedback supports only one refinement, return exactly one path.

Critical rule:

- Never create multiple paths that are only superficial rewrites of the same reasoning.
- Each refinement path must represent a meaningfully distinct clinician-style reasoning emphasis or framing.
- If the feedback does not clearly support multiple refinements, return a single path.

What makes a strong reasoning refinement:

- The reasoning explicitly links the context and question to the clinical decision.
- The reasoning explains why the important details matter clinically.
- The reasoning reflects prioritization, trajectory, safety, alternatives, and decision relevance.
- The reasoning is concise but substantive.
- The reasoning clearly reflects the clinician feedback.
- The reasoning sounds like real ICU chart-review reasoning rather than generic commentary.

Current question:

```
{question}
```

Current context:

```
{context}
```

Current reasoning:

```
{reasoning}
```

Clinician feedback:

```
{feedback_text}
```

Return strict JSON only with exactly this schema:

```
{
  "explanation": "concise clinician-style explanation of what is missing or weak in the
     current reasoning, written as a clean synthesis of the clinician feedback",
  "variant_strategy": "single_variant" | "multi_variant",
  "refinement_paths": [
    {
      "path_id": 1,
      "scenario_label": "short label for this reasoning refinement path",
      "clinical_rationale": "brief clinician-style rationale for why this is a distinct
     reasoning direction or emphasis",
      "improved_reasoning": "improved clinician-style reasoning paragraph",
      "specific_feedback_addressed": ["specific issue 1", "specific issue 2"]
    }
  ],
  "audit": {
    "main_issues_addressed": ["issue 1", "issue 2"],
    "preserved_original_intent": true,
    "used_feedback_count": 0
  }
}
```

Rules:

- Output valid JSON only.
- Do not wrap in markdown fences.
- `explanation` should be concise, specific, and faithful to the clinician feedback.
- `explanation` should read like brief clinician reasoning, not robotic meta-commentary.
- `explanation` should state why the current reasoning is insufficient for the decision, not just that it is "too vague".
- Do not include information in `explanation` that was not present in the clinician feedback.
- `refinement_paths` must contain at least one item.
- Each `clinical_rationale` and `improved_reasoning` must be a non-empty string.
- `improved_reasoning` must remain aligned with the provided question and context.
- `improved_reasoning` must be one coherent paragraph in natural language.
- `improved_reasoning` must not invent unsupported patient details.
- If feedback is weak or partial, make only minimal justified edits and return a single path.
- Usually return one to four refinement paths maximum.

Prompt 5

Prompt for selecting relevant variable-taxonomy sections using clinician feedback on ICU-REACT test set.

You are an ICU clinician and annotation reviewer. Your job in this step is only to decide which parts of the variable taxonomy are needed to:

1. cover the original approved variables, and
2. extract any additional variables implied or explicitly requested in the expert feedback.

You will be given:

- original variables that were already approved,
- expert feedback describing what clinicians want added or covered,
- a taxonomy index containing categories and their subcategories, but no variables yet.

Task:

- Select only the categories and subcategories that are relevant to:
  1. mapping the original variables into the taxonomy, and
  2. capturing clinician-requested variables from expert feedback.
- Be conservative: if a variable or request could plausibly belong to a section, include it.
- Do not select irrelevant sections.

Original variables approved by at least one annotator:

```
{variables_str}
```

Expert feedback:

```
{feedback_text}
```

Taxonomy index. Choose from here:

```
{taxonomy_index}
```

Return only the following JSON, with no text outside JSON:

```
{
  "selected_sections": [
    {
      "category": "CategoryName",
      "subcategories": ["SubcatA", "SubcatB"]
    }
  ]
}
```

Constraints:

- Category and subcategory names must match the taxonomy index exactly.
- Only include subcategories that belong to the specified category.
- If nothing applies, which should be rare, return:

```
{
  "selected_sections": []
}
```

Prompt 6

Prompt for selecting refined variables from chosen taxonomy sections on ICU-REACT test set.

You are an ICU clinician and annotation reviewer. Your job in this step is to map the original approved variables to the taxonomy variables provided, when needed, and to select any additional variables requested by expert feedback, constrained to variables by section.
You will be given:

- original selected variables that were already approved,
- expert feedback,
- variables by section: a list of variables grouped under category and subcategory.

The variables by section list is already filtered to only the sections selected in Step A.

Protocol:

1. **Map original variables to taxonomy, if needed.**
   - For each item in original variables, try to find an exact match in variables by section.
   - If an original variable does not appear exactly, choose the closest proxy from variables by section and record it in `original_proxy_mapping`.
   - The mapped original set is called `original_mapped`, using exact matches and/or proxies.
2. **Extract clinician requests.**
   - From expert feedback, list the explicit variables or concepts clinicians are asking for using short clinical labels.
3. **Add needed variables from the provided section lists.**
   - Choose variables only from variables by section and match names exactly.
   - If a requested concept lacks an exact match, choose the closest proxy from variables by section and record it in `feedback_proxy_mapping`.
   - Be conservative and clinically focused; avoid redundancy and near-duplicates.
4. **Build outputs.**
   - `added_from_feedback`: items added for feedback that were not in `original_mapped`.
   - `refined_variables`: `original_mapped` ∪ `added_from_feedback`, deduplicated.
5. **Coverage check.**
   - `all_original_mapped` is true only if every original variable is covered either by an exact match in `original_mapped` or via `original_proxy_mapping`.
   - `all_requested_included` is true only if every requested concept is covered either by an exact variable in `refined_variables` or via `feedback_proxy_mapping`.
   - `missing_original`: original variables not covered, empty if none.
   - `missing_requested`: requested concepts not covered, empty if none.

Original variables selected, approved by at least one annotator:

```
{variables_str}
```

Expert feedback:

```
{feedback_text}
```

Variables by section. Select only from these and match names exactly:

```
{variables_by_section}
```

Return only the following JSON, with no text outside JSON:

```
{
  "audit": {
    "requested_from_feedback": ["..."],
    "original_selected": ["..."],
    "coverage_check": {
      "all_original_mapped": true or false,
      "missing_original": ["..."],
      "all_requested_included": true or false,
      "missing_requested": ["..."]
    },
    "original_proxy_mapping": [
      {"concept": "original variable label", "proxy": "ExactVariableName"}
    ],
    "feedback_proxy_mapping": [
      {"concept": "requested label", "proxy": "ExactVariableName"}
    ],
    "omitted_with_reason": [
      {"concept": "requested label", "reason": "brief clinical justification"}
    ]
  },
  "original_mapped": ["Var1", "Var2", "..."],
  "added_from_feedback": ["VarA", "VarB"],
  "refined_variables": ["Var1", "Var2", "..."]
}
```

Constraints:

- Every item in `original_mapped`, `added_from_feedback`, and `refined_variables` must be drawn exactly from variables by section or included via the corresponding proxy mappings.
- Do not invent new variables or rename variables.
- If nothing new is required from feedback, return `added_from_feedback: []` and keep `refined_variables` equal to `original_mapped`.

Prompt 7

Prompt for generating refined reasoning from clinician-selected variables and feedback on ICU-REACT test set.

You are an ICU clinician. Refine the reasoning paragraph for the decision-making question using the provided variables grouped by category and sub-category. Do not change the variable list. Output only valid JSON.

Goal:

- Produce a single natural-language paragraph explaining how the provided variables support answering the question.

- Align the reasoning with the expert feedback.

Introduction:
Start the paragraph with one introductory sentence that:

1. briefly re-states the patient context and the decision at hand, and
2. provides a brief overview of the key types of data that need to be reviewed, such as laboratory results, physiology and vital signs, diagnoses, imaging, medications, or fluid balance, without listing every variable.

Length and style requirements:

- Write eight to twelve sentences total, including the introductory sentence.
- Target approximately 220–350 words.
- Keep the writing fluent and clinical.
- Avoid repetitive phrasing such as "X matters because ...".
- Do not sound like a taxonomy dump or checklist.

How to use the taxonomy grouping:

- Use categories and sub-categories as light structural cues.
- Do not write sentences of the form:

```
<Category> <Sub-category> helps with ...
<Sub-category> helps ...
```

Bad example to avoid:

```
Laboratory Measurements Comprehensive Metabolic Panel helps with ...
```

Instead, follow this pattern:

1. When first introducing a category, write one sentence explaining why that category matters for the decision. This should be a general rationale and should not yet list variables.
   Examples:

   ```
   Laboratory Measurements are essential to quantify severity and
   immediacy of physiologic derangements.

   Physiology helps assess stability, trajectory,
   and response to interventions.
   ```

2. For each sub-category within that category, write one subsequent sentence that:
   - starts with the sub-category name, or a natural transition referencing it,
   - lists the variables as a compact cluster using comma-separated formatting,
   - provides one concise rationale for the entire cluster.

   Example:

   ```
   Comprehensive Metabolic Panel (BUN, creatinine,
   potassium, bicarbonate) frames renal clearance
   and immediate electrolyte and acid-base threats.
   ```

Additional writing guidance:

- Vary transitions across categories.
- Avoid repeatedly starting sentences with the same structure.
- Examples include:
  - Laboratory Measurements...
  - Physiology...
  - On the diagnosis side...
  - Imaging...
  - Medication context...

  - Fluid balance...
- Do not stack category and sub-category names back-to-back in the same clause unless grammatically necessary.
- When referencing categories or sub-categories, use their full names exactly as provided in VARIABLES.
- Match spelling and capitalization exactly.
- Avoid abbreviations unless the full name is written first.
- Only single out an individual variable with extra detail if it is central to the decision, and limit this to at most one or two variables.

Coverage requirement:
Every variable must be addressed by appearing explicitly somewhere in the paragraph:

- either as part of its sub-category cluster,
- or as an individually highlighted variable.

If the variable list is long, coverage should primarily occur through clustered mentions rather than repeated references.

Hard rules:

- Do not invent patient values.
- Do not make a final recommendation.
- Do not include meta-commentary about the task.
- If the current reasoning is incomplete or inconsistent, rewrite it into a complete and internally consistent paragraph suitable for fine-tuning.

Question:

`{question}`

Context:

`{context}`

Variables (grouped by Category/Sub-category; do not modify):

`{refined_variables}`

Current reasoning:

`{reasoning_text}`

Expert feedback:

`{feedback_text}`

Return only the following JSON, with no text outside JSON:

```
{
  "audit": {
    "coverage_check": {
      "all_variables_addressed": true or false,
      "missing_in_reasoning": ["..."]
    }
  },
  "refined_reasoning":
    "One paragraph (8-12 sentences) starting with an
     introductory sentence that re-states the context
     and decision and summarizes what data must be
     reviewed, then explains how the grouped variables
     support the decision, aligned with expert feedback."
}
```

Constraints:

- Do not add, remove, rename, or reorder variables.
- Use exactly the variables provided in VARIABLES.

Prompt 8

Prompt for creating case-specific LLM-judge rubrics on ICU-REACT test set.

You are helping build an ICU decision-making benchmark called ICU-REACT.
For each single evaluation sample, you will design an explicit rubric that can be used by another LLM-judge to grade candidate responses.

Critical constraint:
The LLM-judge will only see:

1. the patient context,
2. the decision-making question,
3. the candidate model response,
4. the rubric JSON you write.

The judge will not have access to the reference answer. Therefore, every rubric criterion must be self-contained, observable from the candidate response, and grounded in the patient context and question. Do not write criteria that require access to hidden ground truth wording.

Each sample consists of:

- Patient context, representing an ICU scenario.
- A decision-making question.
- A reference answer written by an expert, visible to you only, which indicates:
  - which categories, subcategories, and variables are relevant,
  - why they are relevant,
  - how they influence the decision,
  - relevant safety considerations.

Your task is to read the sample and create a case-specific rubric that evaluates the quality and correctness of the response's reasoning.

Important goal:
The rubric should primarily evaluate whether the candidate response gives the correct clinical reason for including categories, subcategories, and variables.
This means:

- Do not reward a response simply for mentioning the right variable.
- Instead, reward whether the response correctly explains:
  - why that variable or category matters,
  - what clinical role it plays,
  - and how it influences the decision.
- A response should score well only if it provides correct, decision-relevant, mechanistic reasoning.

Target behavior of the rubric:

- A strong response explains the correct reason that each important factor matters.
- A weak response may mention relevant factors but give vague, generic, incomplete, or incorrect reasons.
- A response should not receive high scores for broad variable listing alone.
- A concise response with correct and well-explained reasoning should outperform a longer but shallow response.
- The rubric should reward explanation quality, not mention count.

Task-faithfulness requirement:

- Candidate responses should not make a specific final recommendation or decision.

- They should focus on explaining the decision factors: categories, subcategories, variables, and how each would influence the decision.
- Therefore, the rubric must evaluate faithfulness to this format: explaining decision factors without recommending a final action.

Example ICU-REACT case and rubric, for guidance only.
Do not reuse the same wording. Adapt the style to the new sample.

Example context:

```
60-year-old male with acute kidney injury and electrolyte disturbances in the ICU.
```

Example question:

```
Should we initiate dialysis for this patient with severe electrolyte imbalances?
```

Example reference answer, visible to you only and not to the judge:

```
In this 60-year-old man in the ICU with acute kidney injury and severe electrolyte
     disturbances, the decision to initiate dialysis depends on integrating laboratory
     derangements, physiologic stability and fluid status, cardiac/imaging evidence of
     electrolyte toxicity or overload, neurologic/physical exam findings, and current
     medication exposures that may worsen kidney function or electrolytes. Medications are
     critical to interpret because recent or ongoing therapies can precipitate acute kidney
     injury, aggravate electrolyte abnormalities, and determine whether medical management
     has been maximized or is limited by toxicity. Fluids (Albumin 25%, Albumin 5%, Lactated
     Ringers, Normal Saline) inform the resuscitation strategy and whether administered
     volume or oncotic support could be contributing to or mitigating overload while kidney
     function is impaired. Diuretics (Bumetanide, Furosemide, Metolazone, Spironolactone)
     clarify attempts to manage volume and potassium balance, and lack of response despite
     these agents supports concern for refractory overload or electrolyte control.
     Antibiotics (Cefepime, Gentamicin, Levofloxacin, Meropenem, Piperacillin-Tazobactam,
     Vancomycin) are particularly relevant as potential nephrotoxins or contributors to acute
      kidney injury and may necessitate dialysis for clearance or to safely continue therapy.
      Antihypertensives (Enalaprilat, Nitroprusside) provide context for renal perfusion and
     hemodynamic tolerance of ultrafiltration in a patient with labile pressures. Laboratory
     Measurements are essential to quantify the severity and immediacy of physiologic
     derangements that dialysis can rapidly correct when conservative therapy is insufficient
     . Comprehensive Metabolic Panel (BUN, Calcium (total), Chloride, Creatinine, Magnesium,
     Phosphorus, Potassium, Sodium) frames the degree of renal failure and the urgency of
     threats such as potassium elevation or progressive azotemia, while also characterizing
     concurrent abnormalities including hypomagnesemia and disordered calcium/phosphate
     handling. Physiology helps assess stability, trajectory, and end-organ impact of
     electrolyte and volume disturbances to judge whether the patient is tolerating ongoing
     medical management. Vitals (Blood pressure diastolic, Blood pressure systolic, Heart
     rate, Mean arterial pressure, Oxygen saturation, Respiratory rate) highlight hemodynamic
      adequacy and respiratory compromise, with declining Oxygen saturation or rising
     Respiratory rate raising concern for fluid-overload physiology that may push toward
     dialysis-based ultrafiltration. Fluidics (Fluid balance, Fluid intake, Insensible losses
     ) and Excretion (Hourly urine output, Urine output) together determine whether the
     patient is accumulating volume despite therapy and whether oliguria suggests limited
     capacity to correct electrolytes without extracorporeal support. Imaging provides
     objective evidence of cardiopulmonary consequences and cardiac electrical instability
     that can make electrolyte derangements immediately dangerous. EKG (EKG Findings, EKG
     Rhythm) identifies arrhythmias or electrocardiographic manifestations consistent with
     severe potassium or magnesium disturbances, while Basic Radiography (Pulmonary edema (X-
     ray)) supports clinically significant overload. Scores and Assessments help determine
     systemic impact and urgency, as Neuro-Cognitive Assessments (Glasgow Coma Scale) can
     signal encephalopathy from uremia or electrolyte shifts and Physical Exam Findings (
     Peripheral edema) corroborate volume overload. Diagnosis anchors interpretation by
     confirming Renal (acute kidney injury) as the substrate for impaired clearance and
     highlighting Metabolic (Hypomagnesemia) as a documented abnormality that, if severe or
     refractory, may accompany other dialysis-relevant electrolyte threats.
```

Example rubric. JSON shape to imitate:

```
{
  "rubric": [
    {
      "id": "R1",
      "dimension": "task_faithfulness",
      "description": "Evaluates whether the response stays focused on explaining decision
      factors and their influence on the dialysis decision, without making a final
      recommendation to initiate or withhold dialysis.",
      "weight": 5,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R2",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why chemistry and
      renal laboratory abnormalities are relevant: they define the severity of kidney
      dysfunction and electrolyte derangement, and help determine whether abnormalities are
      severe, refractory, or dangerous enough that dialysis becomes more relevant.",
      "weight": 4,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R3",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why diuretics are
      relevant to the dialysis decision: they reflect attempts to manage volume status and
      potassium balance, and lack of response despite these agents increases concern for
      refractory volume overload or inability to control electrolytes with medical therapy
      alone.",
      "weight": 4,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R4",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why fluid
      administration is relevant: fluids such as albumin or crystalloids inform the
      resuscitation strategy and help determine whether administered volume may be
      contributing to, worsening, or attempting to correct hemodynamic or volume problems
      while renal clearance is impaired.",
      "weight": 3,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R5",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why nephrotoxic or
      renally relevant medications, especially antibiotics, matter: they may worsen acute
      kidney injury, aggravate electrolyte disturbances, or influence whether dialysis is
      needed to support clearance or permit ongoing treatment safely.",
      "weight": 4,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R6",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why urine output and
      fluid balance matter: reduced urine output or ongoing positive balance suggests limited
      ability to clear volume and electrolytes without extracorporeal support, increasing
      concern for refractory overload or failed medical management.",
      "weight": 4,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R7",
      "dimension": "reasoning_correctness",
```

```
      "description": "Evaluates whether the response correctly explains why vitals and
      physiologic status matter: blood pressure, MAP, oxygenation, respiratory rate, and
      related instability help assess both the consequences of overload/electrolyte
      derangement and the patient's likely tolerance of dialysis or ultrafiltration.",
      "weight": 3,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R8",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why EKG findings are
      relevant: they provide objective evidence of cardiac electrical instability from
      electrolyte abnormalities, especially dangerous potassium or magnesium disturbances,
      which increases concern for urgent dialysis-relevant complications.",
      "weight": 4,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R9",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why imaging findings
      such as pulmonary edema matter: they provide objective evidence of clinically
      significant volume overload, supporting concern that fluid cannot be adequately managed
      without dialysis-based ultrafiltration.",
      "weight": 3,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R10",
      "dimension": "reasoning_correctness",
      "description": "Evaluates whether the response correctly explains why neurologic or
      exam findings such as reduced GCS, encephalopathy, or peripheral edema matter: they
      reflect systemic consequences of uremia, electrolyte disturbance, or volume overload
      that increase concern for dialysis-relevant deterioration.",
      "weight": 3,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R11",
      "dimension": "reasoning_synthesis",
      "description": "Evaluates whether the response correctly synthesizes multiple
      categories into a coherent clinical explanation, showing how labs, medications, fluids,
      output, vitals, imaging, and exam findings jointly shape the dialysis decision factors
      rather than functioning as isolated observations.",
      "weight": 5,
      "type": "0_1_2_3_scale"
    },
    {
      "id": "R12",
      "dimension": "safety",
      "description": "Evaluates whether the response's explanatory reasoning is clinically
      safe: it should not ignore major dangerous factors, should not assign clearly incorrect
      meanings to variables, and should not rely heavily on weak or irrelevant rationale.",
      "weight": 5,
      "type": "0_1_2_3_scale"
    }
  ]
}
```

This example is only to show the level of specificity and style. For the new sample below, you must design a new rubric with different, case-appropriate criteria.

Now design a rubric for the following sample.

Sample ID:

```
{sample_id}
```

Patient context:

`{context}`

Decision-making question:

`{question}`

Reference answer, expert ground truth, visible to you only and not to the judge:

`{reference_answer}`

General guidelines for your rubric:

1. **The rubric must be specific to this sample, not generic.**
2. **Judge-visibility constraint.**
   The judge will not see the reference answer. Do not write criteria like "mentions the variables in the reference answer." Each criterion must stand on its own and be checkable from the candidate response, using only what is clinically justified by the context and question.
3. **Primary goal: reasoning correctness.**
   The rubric must mainly evaluate whether the response explains the correct reason that a category, subcategory, or variable is relevant. The main target is not coverage by itself, but correctness of explanation. Prefer criteria of the form:

   ```
   factor -> why it matters -> how it influences the decision
   ```

   Good criteria should evaluate whether the response captures the clinical role of the factor, the correct mechanism or interpretation, and the correct decision implication.
4. **Do not over-reward mention count.**
   Do not create criteria that mainly reward listing many factors. Mentioning a factor should only help when the response correctly explains why it matters and how it affects the decision. A vague mention, boilerplate mention, or shallow mention should score low.
5. **Use specific reasoning-correctness criteria.**
   Create criteria for the most important categories, subcategories, or variables from the reference answer. For each important factor, ask whether the response correctly explains why it is relevant, what it clarifies, what concern it raises or lowers, or how it changes interpretation of the decision.
   Example pattern:

   ```
   Evaluates whether the response correctly explains that [factor] matters because [clinical
        role], and that it influences the decision by [decision implication].
   ```

6. **Use synthesis criteria in addition to local correctness criteria.**
   Include at least one criterion evaluating whether the response integrates multiple factors into a coherent clinical explanation rather than listing them as isolated facts. This criterion should reward cross-category reasoning and good synthesis.
7. **Task faithfulness.**
   Include task-faithfulness criteria that evaluate whether the response remains in decision-factor analysis mode. The response should explain what factors matter and how they influence the decision. The response should not make a final recommendation such as initiating, withholding, or choosing a definitive action. Penalize recommendation-making or drifting into final action selection.
8. **Safety.**
   Include safety-oriented evaluation of the explanatory reasoning. Safety here means the response should not:
   - ignore major dangerous factors,
   - assign clearly incorrect meanings to important variables,
   - rely on obviously weak or irrelevant reasoning,
   - or present clinically misleading interpretations.

   Safety should be evaluated in terms of reasoning quality, not final management choice.

9. **Criterion granularity.**
   Criteria should be atomic and checkable. Each criterion should evaluate one clear reasoning behavior. Avoid bundling multiple unrelated expectations into one criterion. However, it is acceptable for a criterion to cover a tightly connected cluster if those elements share the same clinical role and decision implication.
10. **Number of criteria.**
    There is no fixed required number of criteria. Include as many criteria as needed to cover the important reasoning components of the sample well. Prefer completeness of important reasoning over forcing an arbitrary count. In most cases, the rubric will likely have roughly eight to fourteen criteria, but this is not a hard rule.
11. **Weighting philosophy.**
    Reasoning correctness should dominate the rubric. High-value criteria should reward correct explanation of important factors. Synthesis, task faithfulness, and safety should also be represented. Do not overuse low-value binary criteria. Prefer giving more weight to clinically central reasoning elements.
12. **Preferred criterion types.**
    Use `0_1_2_3_scale` for most reasoning-correctness criteria. This scale should reflect quality of explanation, for example:
    - 0 = absent, incorrect, or clinically misleading.
    - 1 = factor is mentioned but reasoning is vague, shallow, incomplete, or partly incorrect.
    - 2 = reasoning is mostly correct but incomplete or insufficiently specific.
    - 3 = reasoning is clearly correct, specific, and decision-relevant.

    Use `0_1_2_scale` when a simpler partial-credit scale is sufficient. Use `present_absent` only for rare must-have anchors, not for most reasoning criteria.
13. Do not invent new patient details.
14. **Do not design the rubric around final recommendation correctness.**
    The response is not supposed to choose the final action. Therefore, do not reward or penalize based on whether the response recommends a specific final decision. Instead, evaluate whether the response correctly explains the decision factors.

Allowed fields:
For each criterion, output:

- `id`: a short string like `R1`, `R2`, etc.
- `dimension`: one of:
  - `reasoning_correctness`
  - `reasoning_synthesis`
  - `task_faithfulness`
  - `safety`
  - `critical_anchor`
- `description`: concise and case-specific, written so a judge without the reference answer can apply it using only the context, question, and response.
- `weight`: integer points from 1 to 5.
- `type`: one of:
  - `present_absent`
  - `0_1_2_scale`
  - `0_1_2_3_scale`

Additional guidance on dimensions:

- `reasoning_correctness`: use for criteria that evaluate whether the response correctly explains why a factor matters and how it influences the decision.
- `reasoning_synthesis`: use for criteria that evaluate integration across multiple factors or domains.
- `task_faithfulness`: use for criteria that evaluate staying in decision-factor explanation mode without making a final recommendation.
- `safety`: use for criteria that evaluate whether the explanatory reasoning is clinically safe.
- `critical_anchor`: use sparingly for rare must-not-miss anchors.

Weighting guidance:

- `reasoning_correctness`: typically 3 to 5.
- `reasoning_synthesis`: typically 4 to 5.
- `task_faithfulness`: typically 4 to 5.
- `safety`: typically 4 to 5.
- `critical_anchor`: typically 1 to 3.

Output format:
Return a single JSON object with the following structure:

```
{
  "rubric": [
    {
      "id": "R1",
      "dimension": "reasoning_correctness",
      "description": "...",
      "weight": 4,
      "type": "0_1_2_3_scale"
    }
  ]
}
```

Rules:

- Do not include explanations outside of this JSON.
- Do not include comments, markdown, or trailing commas.
- Make sure the JSON is syntactically valid.
- Ensure every criterion is case-specific.
- Ensure the rubric emphasizes reasoning correctness over mere coverage.

S13.2 Dataset augmentation prompts

Prompt 9

Prompt for generating ICU-REACT context–question pairs for dataset augmentation.

You are an ICU clinical authoring assistant. Your task is to generate `{n}` new context–question pairs to augment a dataset for instruction tuning.

General constraints, applied to every generated pair:

- Language: `{language}`.
- Do not include any specific numeric values for laboratory results, vital signs, ventilator settings, medication doses, or flows. Age is allowed.
- Keep each pair ICU-relevant, clinically plausible, and free of protected health information, including names, exact dates, addresses, and medical record numbers.
- Each pair must be distinct and should plausibly lead to different reasoning or variable selections during annotation.
- The context should be one to two sentences and include key clinical features, etiologies, or constraints.
- Age is allowed, but avoid numeric measurements.
- The context should follow this format:

```
[age][gender] patient with [past_medical_history] admitted to
the ICU due to [surgery_or_current_complication] is developing
[critical_condition]
```

- The question should be one sentence, action-oriented, and decision-focused.
- The question may focus on diagnostic, therapeutic, monitoring, escalation, or modality-choice decisions.
- Avoid wording that implies numeric levels. For example, say "concern for oxygenation" instead of "high PEEP".
- Each pair must be categorized under exactly one of the ICU topics listed below, based on the primary clinical issue and decision focus.

When creating the context–question pairs, try to ground them in the ICU topics below. For each topic, ideas are suggested but may be adapted to the specific patient story.

- **Topic 1**
  - **Example clinical scenarios:** ...
  - **Typical ICU presentations to highlight:** ...
  - **Decisions the question can target:** ...
  - **Possible modifiers or constraints:** ...

  ...
- **Topic n**
  - **Example clinical scenarios:** ...
  - **Typical ICU presentations to highlight:** ...
  - **Decisions the question can target:** ...
  - **Possible modifiers or constraints:** ...

Few-shot examples for style and content guidance. Do not copy blindly:

```
{few_shot_examples}
```

Output only a JSON object with this exact schema, with no extra keys and no prose outside JSON:

```
{
  "output": [
    {
      "context": "Generated patient context here",
      "question": "Generated decision-making question here",
      "topic": "One of the ICU topics listed above"
    }
  ]
}
```

Prompt 10

Prompt for generating initial ICU-REACT reasoning from context–question pairs.

You are an ICU clinical authoring assistant. You will receive an ICU patient context and a decision-making question.

Your task:

1. Write one initial reasoning paragraph that explains what information would matter to answer the question.
2. The reasoning should be clinically plausible, relevant, and coherent, but it does not need to be perfect.
3. The reasoning should feel like an early draft that could later be improved by expert refinement.

Initial reasoning guidelines:

- Language: English.
- Write exactly one paragraph.

- Do not output bullets, numbered lists, section headers, or JSON-like text inside the reasoning.
- Keep the reasoning decision-focused and clinically grounded.
- Emphasize what information, physiology, trajectories, risks, or alternative explanations matter.
- Avoid numeric specifics for laboratory results, vital signs, ventilator settings, medication doses, or flows.
- Age is allowed if already present in the context.
- Do not mention that the reasoning is incomplete or intentionally imperfect.
- Do not copy the question verbatim as the answer.

Few-shot examples for style guidance. Do not copy blindly:

```
{few_shot_examples}
```

Input:

Context:

```
{context}
```

Question:

```
{question}
```

Output only a JSON object with this exact schema, with no extra keys and no prose outside JSON:

```
{
  "output": [
    {
      "question": "The input question here",
      "context": "The input context here",
      "initial_reasoning": "Initial clinical reasoning here"
    }
  ]
}
```

Prompt 11

Prompt for generating refined ICU-REACT reasoning from an initial reasoning draft.

You are an ICU clinical educator. You will receive:

- an ICU patient context,
- a decision-making question,
- an initial reasoning draft.

Your task:

1. Analyze the initial reasoning and identify how it should be improved.
2. Write a concise explanation of the main refinement opportunity.
3. Write one improved reasoning paragraph.

Refinement goals:

- Preserve the original clinical intent unless it is clearly misguided.
- Improve prioritization, clinical framing, explanatory quality, and relevance.
- Strengthen links between the patient's presentation and the decision being asked.

- Clarify the most decision-relevant physiology, risks, trajectory, alternatives, and monitoring needs.
- If appropriate, include important alternative explanations or parallel checks that should be considered.
- Keep the reasoning ICU-relevant and clinically plausible.

Output guidelines:

- Language: English.
- The explanation should be concise but specific.
- The improved reasoning must be exactly one paragraph.
- Do not output bullets, numbered lists, or extra commentary outside the JSON schema.
- Avoid numeric specifics for laboratory results, vital signs, ventilator settings, medication doses, or flows.
- Age is allowed if already present in the context.
- Do not generate multiple variants or branches.
- Do not use the terms "refinement path" or "path".
- Keep the improved reasoning natural-language only.

Few-shot examples for style and refinement strategy guidance:

```
{few_shot_examples}
```

Input:

Context:

```
{context}
```

Question:

```
{question}
```

Initial reasoning:

```
{initial_reasoning}
```

Output only a JSON object with this exact schema, with no extra keys and no prose outside JSON:

```
{
  "output": [
    {
      "question": "The input question here",
      "context": "The input context here",
      "initial_reasoning": "The input initial reasoning here",
      "reasoning_refine_explanation": "Explanation of how the reasoning should be improved",
      "improved_reasoning": "Improved reasoning here"
    }
  ]
}
```

S13.3 Training prompts

Prompt 12

Prompt for ICU-REACT reasoning refinement training.

You are an ICU clinician. Your task is to review an initial clinical reasoning draft for an ICU patient context and decision-making question, explain how the reasoning should be improved, and then write an improved reasoning paragraph.
Output the explanation followed by the improved reasoning in the exact natural-language format requested. Do not output JSON, bullet lists, or extra commentary.

Patient context:

```
{context}
```

Decision-making question:

```
{question}
```

Initial reasoning:

```
{initial_reasoning}
```

Writing requirements:

- First explain what is missing, overemphasized, underemphasized, or clinically misprioritized in the initial reasoning.
- Then write one improved reasoning paragraph.
- Keep the explanation concise but specific.
- Keep the improved reasoning clinically coherent, ICU-relevant, and decision-focused.
- Do not output JSON.

Output exactly in this format:

```
Reasoning refinement explanation:
<explanation>

Improved reasoning:
<one coherent paragraph>
```

Prompt 13

Prompt for ICU-REACT context–question refinement training.

You are an ICU clinician. Your task is to take an underspecified ICU patient context and decision-making question and rewrite it into multiple clinically distinct scenario-specific formulations when appropriate.
Output a shared clinical explanation followed by all clinically distinct scenarios in the exact natural-language format requested. Do not output JSON, bullet lists, or extra commentary.

Initial patient context:

```
{initial_context}
```

Initial decision question:

```
{initial_question}
```

Writing requirements:

- First explain why the initial context–question pair is too underspecified for ICU decision-making.
- Then provide all clinically distinct scenario-specific formulations that should be considered.
- Each clinical scenario must include:
  - scenario label,
  - clinical rationale,
  - scenario-specific question,
  - scenario-specific context.
- Keep everything clinically coherent and ICU-relevant.
- Do not output JSON.

Output exactly in this format:

```
Clinical explanation:
<shared explanation>

Clinical scenario 1
Scenario label: <label>
Clinical rationale: <rationale>
Scenario-specific question: <question>
Scenario-specific context: <context>

Clinical scenario 2
Scenario label: <label>
Clinical rationale: <rationale>
Scenario-specific question: <question>
Scenario-specific context: <context>

(continue for all clinically distinct scenarios)
```

Prompt 14

Prompt for ICU-REACT context generation training.

You are an ICU clinician. Your task is to write one realistic patient context for which the given decision-making question would be clinically relevant.
Output one patient context only. Do not add explanations, bullet points, or JSON. Be specific, clinically grounded, and include enough detail to make the question applicable.

Decision question:

```
{question}
```

Writing requirements:

- Output exactly one patient context, with no extra text.
- The context must make the decision question clinically applicable.
- Include enough ICU-relevant detail, such as organ support, physiology, trajectory, key comorbidities, or acute problem, to justify why this question would be asked.
- Do not answer the question.
- Do not invent unnecessary details beyond what is needed to make the question realistic and clinically coherent.

Prompt 15

Prompt for ICU-REACT question generation training.

You are an ICU clinician. Your task is to write one decision-making question that would guide immediate clinical actions for the patient described.
Output one question only. Do not add explanations, bullet points, or JSON. Be specific, clinically grounded, and focused on a real ICU decision.

Patient context:

```
{context}
```

Writing requirements:

- Output exactly one question, with no extra text.
- The question must be a decision-making question, such as what to do, whether to do a specific action, or what is driving a clinical change.
- Avoid vague questions such as "What is going on?".
- Be specific to likely ICU decisions, such as fluids, vasopressors, ventilation, antibiotics, sedation, or diagnostics.
- Do not invent data that is not in the context.

Prompt 16

Prompt for ICU-REACT variable selection reasoning training.

You are an ICU clinician. Write one coherent paragraph explaining which clinical data are most relevant to the decision and why. Do not use bullet points. Do not output JSON.
Output plain text only. Be concise, fluent, and clinically grounded.

Patient context:

```
{context}
```

Decision-making question:

```
{question}
```

S13.4 Inference prompts

Prompt 17

Prompt for inference on ICU-REACT test set.

You are an ICU clinician. Write one coherent paragraph explaining which of the provided variables are most relevant to the decision and why.
You may reference only a clinically meaningful subset of the provided variables. Do not try to cover all of them. Do not use bullet points. Do not output JSON. Output plain text only. Be concise.

Patient context:

```
{context}
```

Decision-making question:

```
{question}
```

Provided variables, grouped by Category/Sub-category. Use only these variable names when referencing data:

`{grouped_variables}`

Writing requirements:

- Output one paragraph only, with no headings, bullets, or JSON.
- Write eight to twelve sentences, approximately 220–350 words.
- Start with one introductory sentence that restates the patient context and decision at hand and briefly summarizes what data need review.
- Focus only on the variables that are clinically relevant to this patient and this decision.
- Do not attempt to mention or justify every provided variable.
- Aim to reference a representative, decision-informing subset across the relevant categories and sub-categories, including severity, trajectory, contraindications, response to therapy, and safety monitoring.
- When first introducing a category, write one sentence explaining why that category matters for this decision. This should be a general rationale without listing variables yet.
- For each sub-category you choose to include, write one sentence that clusters the most relevant variables using comma-separated formatting and gives one concise rationale for the cluster.
- It is acceptable to omit entire categories or sub-categories if they are not decision-informing for this case.
- Include a final closing sentence that briefly summarizes how the selected data will support making the decision, without making a recommendation.
- Do not invent patient values or make a final recommendation.
- Every variable you mention must match exactly a name from the provided variables list.

Prompt 18

Prompt for inference on SCT-Bench test set.

You are taking a Script Concordance Test.
You will receive a single clinical case consisting of a scenario, a hypothesis, and additional information. Ignore any examples in the prompt; they are already answered. For this case, respond with exactly one line starting with `Rating:` followed by one of `{-2, -1, 0, +1, +2}`, then a brief explanation.
Do not ask for more scenarios or start a conversation. Do not describe the test instructions back to me.

Guideline:

`{guideline}`

Worked examples, read-only.
The following examples are already scored. Do not score them again.

Example with rating -2:

`{example_-2}`

Example with rating -1:

`{example_-1}`

Example with rating 0:

`{example_0}`

Example with rating +1:

{example_+1}

Example with rating +2:

{example_+2}

New SCT item to score.
Now, you will be given one new scenario, hypothesis, and additional information. Ignore the previous examples. Your job is to rate only this new item.

Scenario:

{{ scenario }}

Hypothesis:

{{ hypothesis }}

Additional information:

{{ additional information }}

Required response format:

Rating: <one of -2, -1, 0, +1, +2> <brief explanation>

Prompt 19

Prompt for decision factors inference on ER-Reason test set.

You are an experienced Emergency Department physician. Your task is only to describe the medical decision factors needed to narrow the differential diagnosis for the patient's current chief complaint. Do not provide a differential diagnosis list, treatment plan, or disposition. Focus only on what additional history, exam findings, tests, imaging, and consultations are needed.

Based on the patient's chief complaint, age, sex, and available clinical information, provide only the following:

Medical decision factors:
List the additional history you would want, important exam findings you would look for, tests or imaging you would order, and any specialist consultations you would consider in order to narrow the diagnosis for the current chief complaint.
Nothing has currently been performed.

Important temporal context:
Everything in the notes is from past encounters. The patient is now presenting with a new complaint.

Chief complaint, current encounter:

{chief}

Age:

{age}

Sex:

{sex}

Patient one-liner summary, if available:

{one_liner}

Previous hospital encounter notes, for context only:

{notes_content}

Prompt 20

Prompt for differential diagnosis inference on ER-Reason test set.

You are an experienced Emergency Department physician. Your task is only to generate an initial differential diagnosis for the patient's current chief complaint.
Do not provide medical decision factors, workup recommendations, treatment plans, or disposition. Focus only on the differential diagnosis and brief reasoning.

Based on the patient's chief complaint, age, sex, and available clinical information, provide only the following:

Differential diagnosis:
List the most relevant differential diagnoses for the current chief complaint, explain briefly why each is being considered, and note which diagnoses are lower on the differential and why.

Important temporal context:
Everything in the notes is from past encounters. The patient is now presenting with a new complaint.

Chief complaint, current encounter:

{chief}

Age:

{age}

Sex:

{sex}

Patient one-liner summary, if available:

{one_liner}

Previous hospital encounter notes, for context only:

{notes_content}

Prompt 21

Prompt for treatment plan inference on ER-Reason test set.

You are an experienced Emergency Department physician. Your task is only to propose the initial treatment plan for the patient's current ED visit based on the available history and physical exam information.
Do not provide a full differential diagnosis list or a separate decision-factors section. Focus only on the most likely working diagnosis, immediate management, additional diagnostic orders needed for treatment planning, and disposition.

Based on the patient's chief complaint, age, sex, and available clinical information, provide only the following:

Treatment plan:
State the most likely working diagnosis, the recommended initial treatment plan, any additional diagnostic tests you would still order as part of management, and the recommended disposition, such as discharge, admission, or ICU. If there is not enough information for a final diagnosis, state what is still needed.

Current ED encounter physical/history, current visit, if available:

`{ed_presentation}`

Important temporal context:
Everything in the notes is from past encounters. The patient is now presenting with a new complaint.

Chief complaint, current encounter:

`{chief}`

Age:

`{age}`

Sex:

`{sex}`

Patient one-liner summary, if available:

`{one_liner}`

Previous hospital encounter notes, for context only:

`{notes_content}`

Prompt 22

Prompt for inference of assessment recommendations on MedRBench test set.

You are a professional doctor.
Please thoroughly examine the patient case summary presented below. Your objective is to perform a detailed diagnostic analysis utilizing all available information. Note that, due to the potentially limited details, the preliminary diagnosis may encompass several possible conditions.
If you determine that the provided data are inadequate for a definitive conclusion, please enumerate any additional diagnostic tests or information that would be necessary. However, if you can deduce a conclusive diagnosis, please proceed to provide it. Too many requests for information are also inappropriate.

Patient case summary:

`{case}`

Guidelines:

- Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.
- Formulate differential diagnoses based on your analysis.

- If the information is not sufficient for a conclusive diagnosis, specify the further tests or details required.
- Always follow the response format in each turn of the dialogue.
- Never change the section headers using the ### format.

Response format:

```
### Clinical Reasoning:
[Provide a concise step-by-step clinical reasoning summary, with each logical step in a
     separate paragraph, using labels such as <step 1>.]
<step 1> Specific reasoning content of this step
<step 2> Specific reasoning content of this step
...
<step n> Specific reasoning content of this step

### Conclusion:
[Give a preliminary conclusion if possible, or summarize the current findings.]

### Additional Information Required:
[Indicate if further information is needed by specifying the required tests or data. If a
     conclusive diagnosis has been made and no additional information is necessary, only
     output "Not required." directly without any other words in this section.]

For example:
Not required.

or

1. Laboratory tests: details
2. Imaging: details
...
```

Prompt 23

Prompt for the GPT-4.1-mini agent to provide requested assessment recommendation information on MedRBench test set.

You are a professional doctor.
You are a medical expert providing guidance to a junior physician on a patient case. The junior physician will ask you for additional diagnostic information based on the patient's case details and any available ancillary test results. Your role is to provide accurate and relevant responses regarding the availability of specific diagnostic information.

Guidelines:

1. You will receive the patient's case information and any relevant ancillary test results.
2. The junior physician will ask questions about additional diagnostic information needed for the case.
3. If relevant ancillary test information is available for the requested diagnostic area, provide the details accurately.
4. If no relevant ancillary test information is available for the requested diagnostic area, simply state: "There is no relevant ancillary test information available for this request."

Patient case:

```
{case}
```

Ancillary test results:

```
{ancillary_test_results}
```

Example interaction:

```
Junior Physician: "Does the patient have any imaging studies like an X-ray or CT scan?"

Your Response:

If there is relevant imaging information available:
"Based on the available ancillary test results, the patient has undergone a chest X-ray which
      shows [specific findings]."

If there is no relevant imaging information available:
"There is no relevant ancillary test information available for this request."
```

Note:
Your responses should be factual and based solely on the provided patient case information and ancillary test results. Avoid speculation or hypotheticals unless explicitly requested.

Prompt 24

Prompt for inference of final diagnosis on MedRBench test set.

You are a professional doctor.
Please make a final diagnosis for the patient in light of the additional information provided below.

Additional information:

```
{additional_information}
```

Guidelines:

- Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.
- Formulate differential diagnoses based on your analysis.
- Always follow the response format in each turn of the dialogue.
- Never change the section headers using the ### format.

Response format:

```
### Clinical Reasoning:
[Provide a concise step-by-step clinical reasoning summary, with each logical step in a
     separate paragraph, using labels such as <step 1>.]
<step 1> Specific reasoning content of this step
<step 2> Specific reasoning content of this step
...
<step n> Specific reasoning content of this step

### Conclusion:
[Directly output the diagnostic result without any other explanation.]
```

Prompt 25

Prompt for inference of treatment plan on MedRBench test set.

You are a professional doctor.
Please carefully study the following patient case summary, conduct a comprehensive and in-depth treatment planning analysis, and clearly provide the selected treatment for the patient.

Patient case summary:

```
{case}
```

Response format:

```
### Clinical Reasoning:
[Provide a concise step-by-step clinical reasoning summary, with each logical step in a
     separate paragraph, using labels such as <step 1>.]
<step 1> Specific reasoning content of this step
<step 2> Specific reasoning content of this step
...
<step n> Specific reasoning content of this step

### Answer:
[Just output the selected treatment for the patient without any other explanation.]
```

Prompt 26

Prompt for inference on diagnostic workflow pipeline for VivaBench test set.

You are a medical diagnostic assistant reviewing one patient case.
Your task is to choose the single best next step in the diagnostic workflow and return it as a JSON object.

Diagnostic workflow rules:

1. At the beginning of the case, you may only choose `history` or `examination`.
2. Before any `investigation` or `imaging` action, you must first provide exactly one `diagnosis_provisional` action.
3. After any `investigation` or `imaging` has been requested, you must not request any further history or physical examination.
4. Select exactly one action per turn.
5. When the available information is sufficient, provide `diagnosis_final`.

Allowed actions:

- `history`: Ask the patient one to two focused history questions in plain language.
- `examination`: Request a focused physical examination and specify the findings you want assessed.
- `diagnosis_provisional`: Give a provisional diagnosis based on the information currently available, before any investigations or imaging.
- `investigation`: Request a non-imaging test, including laboratory tests and bedside or special tests such as ECG, EEG, or pulmonary function testing. When relevant, specify specimen type.
- `imaging`: Request an imaging study performed by radiology or equivalent imaging services, such as X-ray, ultrasound, CT, MRI, PET, or VQ scan. Specify modality and anatomical region.
- `diagnosis_final`: Give the final diagnosis after completing the evaluation.

Important workflow enforcement:

- Do not choose `investigation` or `imaging` until a `diagnosis_provisional` action has already been given earlier in the conversation.
- On the first turn, do not choose `investigation`, `imaging`, or `diagnosis_final`.
- If you have gathered enough information to suspect a diagnosis but have not yet ordered tests, the next step should usually be `diagnosis_provisional`.

Diagnosis formatting rules for both provisional and final diagnosis:

- You may include up to five diagnoses if there is uncertainty or more than one active problem.
- For each diagnosis, provide:
  - `condition`: free-text condition label,

  - `icd_10_name`: ICD-10 diagnostic name,
  - `icd_10`: ICD-10 code,
  - `confidence`: number from 0.0 to 1.0.
- Confidence scores do not need to sum to 1.0.
- For diagnosis actions, the `query` field should contain a JSON list in the following format:

```
[
  {
    "condition": "free text name of the condition",
    "icd_10_name": "icd 10 name of the condition",
    "icd_10": "icd code of the condition",
    "confidence": score
  }
]
```

Return one JSON object with this structure:

```
{
  "reasoning": "brief justification for the next step",
  "action": "one allowed action",
  "query": "specific request, or diagnosis list when action is a diagnosis action"
}
```

Initial patient input format:

```
Patient: {age}-year-old {gender}

History:
{history_text}
```

Prompt 27

Prompt for JSON parse-retry prompt for invalid diagnostic outputs.

Your previous response could not be parsed. Please return exactly one valid JSON object with this structure:

```
{
  "reasoning": "brief justification for the action",
  "action": "chosen action",
  "query": "specific request"
}
```

Previous response:

```
{previous_response}
```

Prompt 28

Prompt for GPT 4.1-mini history information-mapping agent.

You are a medical information-mapping assistant.

Task:

- Read the user's request about symptoms or medical history.

- Match each requested item only to keys that are present in the provided available-keys list.
- For symptom requests, include any specific characteristics requested when available.

Rules:

- Use only keys from the provided available-keys list.
- Do not invent new matched keys.
- If a request does not correspond to any available key, place it in `unmatched`.

Input:

```
Chief complaint: {chief_complaint}
User request: {query}
Available keys: {keys}
```

Return one JSON object in this format:

```
{
  "matched": [
    {
      "query": "relevant phrase from the request",
      "key": "matching key from available keys",
      "addit": ["optional list of requested symptom characteristics"]
    }
  ],
  "unmatched": [
    {
      "query": "phrase that could not be matched",
      "key": "suggested standardized key"
    }
  ]
}
```

Prompt 29

Prompt for GPT 4.1-mini physical examination information-mapping agent.

You are a medical information-mapping assistant for physical examination findings.

Task:

- Read the user request.
- Match each requested physical examination item only to keys in the provided available-keys list.
- If part of the request cannot be matched, place it in `unmatched`.

Input:

```
User request: {query}
Available keys: {keys}
```

Return one JSON object in this format:

```
{
  "matched": [
    {
      "query": "relevant phrase from the request",
      "key": "matching key from available keys"
    }
  ],
```

```
  "unmatched": [
    {
      "query": "unmatched phrase from the request",
      "key": "suggested standardized key"
    }
  ]
}
```

Return only requested items. Do not include extra information.

Prompt 30

Prompt for GPT 4.1-mini laboratory and non-imaging information-mapping agent.

You are a medical information-mapping assistant for laboratory and non-imaging investigations.

Task:

- Read the user request.
- Match each requested investigation only to keys in the provided available-items list.
- If part of the request cannot be matched, place it in `unmatched`.

Input:

```
User request: {query}
Available items: {items}
```

Return one JSON object in this format:

```
{
  "matched": [
    {
      "query": "relevant phrase from the request",
      "key": "matching key from available items"
    }
  ],
  "unmatched": [
    {
      "query": "unmatched phrase from the request",
      "key": "suggested standardized key"
    }
  ]
}
```

Return keys only, not values.

Prompt 31

Prompt for GPT 4.1-mini imaging information-mapping agent.

You are a medical information-mapping assistant for imaging studies.

Task:

- Read the user request.
- Match each requested imaging study only to keys in the provided available-keys list.
- If part of the request cannot be matched, place it in `unmatched`.

Input:

```
User request: {query}
Available keys: {keys}
```

Return one JSON object in this format:

```
{
  "matched": [
    {
      "query": "relevant phrase from the request",
      "key": "matching key from available keys"
    }
  ],
  "unmatched": [
    {
      "query": "unmatched phrase from the request",
      "key": "suggested standardized key"
    }
  ]
}
```

Prompt 32

Prompt for MedQA multiple-choice inference.

You are a helpful medical assistant.
You are a clinician. Read the question and options and pick the single best answer.
First, briefly reason through the clinical information and rule out incorrect options. Then, on a new line at the end, output only the letter A–E.

Question:

```
{question}
```

Options:

```
{options_block}
```

Required final answer format:

```
<single letter A-E>
```

Prompt 33

Prompt for MedMCQA multiple-choice inference.

You are a helpful medical assistant.
You are a clinician. Read the question and options and pick the single best answer.
First, briefly reason through the clinical information and rule out incorrect options. Then, on a new line at the end, output only the letter A–D.

Question:

```
{question}
```

Options:

```
{options_block}
```

Required final answer format:

```
<single letter A-D>
```

Prompt 34

Prompt for MedXpertQA multiple-choice inference.

You are a helpful medical assistant.
You are a clinician answering multiple-choice questions.
First, briefly reason through the clinical information using clinical reasoning, pathophysiology, and guidelines to narrow down the options. Then, on a new line at the end, output only the single best letter A–J.

Question:

```
{question}
```

Options:

```
{options_block}
```

Required final answer format:

```
<single letter A-J>
```

Prompt 35

Prompt for MMLU medical multiple-choice inference.

You are a helpful medical assistant.
You are a clinician. Read the question and options and pick the single best answer.
First, briefly reason through the clinical or biological information and rule out incorrect options. Then, on a new line at the end, output only the letter A–D.

Question:

```
{question}
```

Options:

```
{options_block}
```

Required final answer format:

```
<single letter A-D>
```

S13.5 Evaluation prompts

Prompt 36

Prompt for rubric-based LLM-Judge evaluation on ICU-REACT test set.

You are an expert ICU clinician evaluating the quality of a model's response using a rubric.

Your job:

- Read the patient context, decision-making question, and model response.
- Apply each rubric criterion to the model response.
- For each criterion, decide how strongly it is satisfied using the `type` field.
- The rubric already encodes what an ideal response should contain; you do not need any other reference.

Patient context:

```
{context}
```

Decision question:

```
{question}
```

Candidate model response to be graded:

```
{model_response}
```

Rubric, as a list of criteria with weights and types:

```
{rubric_json}
```

Scoring rules:
For each rubric entry, the `type` tells you how to set `value`.

1. `type = "present_absent"`
   - `value` must be 0 or 1.
   - `value = 1` if the model response clearly satisfies the criterion.
   - `value = 0` if it does not.
2. `type = "0_1_2_scale"`
   - `value` must be 0, 1, or 2.
   - 0 = not satisfied at all.
   - 1 = partially satisfied.
   - 2 = clearly and fully satisfied.
3. `type = "penalty_if_present"`
   - `value` must be 0 or 1.
   - `value = 1` only if the problematic behavior is present in the model response.
   - `value = 0` if the problematic behavior is absent.

Do not try to compute total points yourself. Just assign the correct `value` for each criterion based on the rubric description and the content of the model response.

Output format, strict JSON:
Return a single JSON object:

```
{
  "criteria_results": [
    {
      "id": "<criterion id, e.g. R1>",
      "type": "<copy the 'type' from the rubric for this criterion>",
```

```
      "value": <integer value as defined above>,
      "evidence": "<a short quote or paraphrase from the model response that justifies your
     decision>"
    }
  ]
}
```

Output JSON strictness:

- Return only valid JSON, with no markdown.
- Strings must not contain unescaped double quotes.
- Do not include comments of any kind, including `//`, `/* */`, or trailing annotations.
- Do not escape apostrophes. Write `patient's`, not `patient\'s`.
- In `evidence`, do not write things like `"PaCO2"` or `"ABG"`. Instead, write `'PaCO2'` and `'ABG'` using single quotes, or escape them as `\"PaCO2\"`.
- Do not include stray newlines, trailing commas, or comments.

Rules:

- Do not include any keys other than `sample_id` and `criteria_results`.
- Do not include explanations outside of this JSON.
- Do not include comments, markdown, or trailing commas.
- Make sure the JSON is syntactically valid and fully parsable.

Prompt 37

Prompt for LLM-Judge evaluation of clinical concept-overlap accuracy scoring against physician reference on ER-Reason test set.

You are a strict clinical evaluator.
You will be given two texts:

- Prediction, generated by the model.
- Reference, corresponding to the physician ground truth.

Your task:
Score the clinical accuracy of the prediction compared to the reference on a 0–100 scale.

Scoring strategy. Follow this order:

1. **Recall, add points:** Identify the key clinical concepts present in the reference and check whether the prediction captures them using equivalent terminology. Synonyms, abbreviations, and alternate names count as correct.
2. **Precision, subtract points:** Penalize the prediction for extra concepts not supported by the reference. Do not penalize harmless generalities or standard workflow steps unless they materially change the meaning.

Core rules:

- Ground the evaluation only in the two provided texts.
- Do not infer beyond the texts.
- Handle negation correctly. Negated concepts should not count as present.
- Treat synonyms and abbreviations as equivalent when they clearly refer to the same concept.

Examples of equivalence, non-exhaustive:

- `cmp` = `comprehensive metabolic panel` = `metabolic panel`
- `electrolyte panel` can match `cmp` or `bmp` if clearly referring to serum electrolytes or a metabolic panel.

- `cbc` = `complete blood count`
- `ua` = `urinalysis`
- `hcg` = `pregnancy test`
- `ct a/p` = `ct abdomen and pelvis`
- `cxr` = `chest x-ray`
- `mi` = `myocardial infarction`
- `sob` = `shortness of breath`

Prediction, model:

`{pred}`

Reference, physician ground truth:

`{ref}`

Scoring rubric:

- Start from `recall_points = 0`.
- Extract five to fifteen key concepts from the reference, focusing on the most clinically important concepts.
- Add points for each key concept that is correctly present in the prediction. Synonyms and abbreviations count.
- If almost all key concepts are matched, `recall_points` should be high, approximately 70–100.
- If about half are matched, `recall_points` should be moderate, approximately 40–70.
- If few are matched, `recall_points` should be low, approximately 0–40.
- Then compute `precision_penalty`.
- Subtract points for each extra concept in the prediction that is not supported by the reference.
- Penalize more for high-impact incorrect extras, such as wrong diagnosis, wrong disposition, or high-risk intervention.
- Penalize less for mild over-inclusion, such as generic laboratories or vital signs, unless it dominates the content.
- `final_score` must equal `clamp(recall_points - precision_penalty, 0, 100)`.

Output only a JSON object with this schema, with no markdown and no extra text:

```
{
  "score": number,
  "recall_points": number,
  "precision_penalty": number,
  "final_score": number,
  "matched_key_concepts": [string, ...],
  "missed_key_concepts": [string, ...],
  "extra_concepts": [string, ...]
}
```

Output constraints:

- `score` should be an integer from 0 to 100.
- `recall_points` should represent the 0–100 contribution before penalties.
- `precision_penalty` should represent the 0–100 deduction for unsupported extras.
- `final_score` must equal `recall_points - precision_penalty`, clipped to the range 0–100.
- Use lowercase for concept strings.
- Use short canonical phrases for concepts.
- Keep each concept list to at most 25 items, prioritizing the most clinically important concepts.
- If either text is empty or non-clinical, return a best-effort score with empty lists.
- Return only the JSON object with the required keys.

Prompt 38

Prompt for GPT 4.1-mini agent for parsing additional diagnostic information requirements on MedRBench test set.

Task overview
You will receive raw text from an auxiliary diagnostic or treatment model describing additional information needed for diagnosis. Your task is to convert that text into a JSON dictionary with a list of individual information-requirement objects.

Core extraction rule
Each JSON object must represent exactly one individual test, examination, imaging study, history question group, or other information-gathering action.

Strict extraction rules

1. Split combined statements into separate objects whenever multiple distinct tests, studies, or examinations are mentioned.
2. Never group multiple distinct tests into one object, even if they share the same category or purpose.
3. The `test_name` field must contain only one individual test, study, examination, or action.
4. If the raw text lists multiple items joined by words such as "and", "or", commas, semicolons, slashes, or parentheses with distinct procedures, split them into separate objects whenever they could reasonably be ordered or requested separately.
5. Keep the original meaning and wording as much as possible, but rewrite only as needed to isolate each individual item.
6. Do not add new tests or infer tests not explicitly mentioned in the raw text.
7. Do not omit any test, examination, imaging study, laboratory study, history inquiry, or specialist evaluation mentioned in the raw text.
8. If multiple items share the same purpose, repeat the same or very similar `info_required` text across multiple objects rather than combining them.
9. If one phrase contains both a broad category and examples, extract the specific examples as separate objects whenever they are explicit.
10. If a phrase contains a bundled examination with subcomponents that are part of a single named examination, keep it as one object only if it is clearly one examination. For example, `Ophthalmological examination (visual acuity, intraocular pressure, fundus examination)` can remain one object because it is one named specialist examination with listed components.
11. If two named imaging studies are listed together, they must be split. For example:

```
"Non-contrast CT (NCCT) scan; Non-contrast MRI (NCMRI) scan"
```

becomes two separate objects.

12. If two named laboratory tests are listed together, they must be split. For example:

```
"C-reactive protein (CRP); Complete blood count (CBC) with differential"
```

becomes two separate objects.

13. For culture or PCR phrasing:
    - If explicitly written as separate possible studies such as `nasal swab for culture` and `throat swab for PCR`, split them.
    - If the wording is ambiguous and refers to a single specimen or test choice, preserve the smallest explicit unit from the text without inventing extra tests.

Field definitions
For each object, output:

- `type`: the major category of the item, such as `Laboratory tests`, `Imaging examinations`, `Medical history inquiries`, or `Specialist examination`.
- `test_name`: one specific individual test, study, examination, or action only.
- `info_required`: the specific purpose, information sought, or reason for requesting that item.

Output requirements

1. Output valid JSON only.
2. Do not include any explanatory text before or after the JSON.
3. Use the following exact JSON format.

```
{
  "items": [
    {
      "type": "Major Category",
      "test_name": "One individual test/study/exam/action",
      "info_required": "Specific purpose or information sought"
    }
  ]
}
```

Examples

Example input:

```
Laboratory tests: C-reactive protein (CRP); Complete blood count (CBC) with differential (WBC
     , neutrophils).
```

Example output:

```
{
  "items": [
    {
      "type": "Laboratory tests",
      "test_name": "C-reactive protein (CRP)",
      "info_required": "Assess inflammatory or infectious activity."
    },
    {
      "type": "Laboratory tests",
      "test_name": "Complete blood count (CBC) with differential",
      "info_required": "Assess leukocytosis, white blood cell count, and neutrophilia."
    }
  ]
}
```

Example input:

```
Imaging: Non-contrast CT (NCCT) scan; Non-contrast MRI (NCMRI) scan.
```

Example output:

```
{
  "items": [
    {
      "type": "Imaging examinations",
      "test_name": "Non-contrast CT (NCCT) scan",
      "info_required": "Evaluate orbital and sinus abnormalities."
    },
    {
      "type": "Imaging examinations",
      "test_name": "Non-contrast MRI (NCMRI) scan",
      "info_required": "Evaluate orbital and sinus abnormalities and intracranial extension."
    }
  ]
}
```

Example input:

```
Ophthalmological examination (visual acuity, intraocular pressure, fundus examination).
```

Example output:

```
{
  "items": [
    {
      "type": "Specialist examination",
      "test_name": "Ophthalmological examination",
      "info_required": "Assess the eye's condition, including visual acuity, intraocular
     pressure, and fundus findings."
    }
  ]
}
```

Raw output text to be organized:

```
{info_required}
```

Prompt 39

Prompt for GPT-4.1 mini information matching agent between prediction and ground truth on MedRBench test set.

Task overview
This task aims to accurately analyze the given categories of medical examination items, specific item names, and their testing purposes to determine whether this information is reflected or covered in the provided reference text.

Task requirements
Given the description to be analyzed, judge whether the medical examination item in the description is the same as, or appears in, one of the examination items in the reference text, or whether the testing purpose or required information in the description is reflected or covered in the reference text.

1. If the specific examination item name or its alias mentioned in the description to be analyzed appears in the reference text, output `Yes`.
2. If the testing purpose or required information mentioned in the description to be analyzed is covered or reflected in the reference text, output `Yes`.
3. If neither the specific examination item name nor the required information content mentioned in the description to be analyzed appears in the reference text, output `No`.

Output requirements
Only output your judgment result on the description to be analyzed, with optional values `Yes` or `No`. Do not output any other content.

Output format

```
[Yes|No]
```

Description to be analyzed:

```
{a_info_step}
```

Reference text:

```
{gt_info}
```

Prompt 40

Prompt for LLM-Judge evaluating diagnostic accuracy against ground truth on MedR-Bench test set.

Task description
You are a professional medical diagnosis evaluation system. You will receive two diagnosis results: one is the diagnosis predicted by the model, and the other is the verified correct diagnosis. Your task is to judge whether the model-predicted diagnosis is correct.

When evaluating, consider the following factors:

1. The same disease may have multiple aliases. For example, `Heart disease` may also be called `Cardiac disease`.
2. There may be diversity in language expression. For example, `heart attack` and `myocardial infarction` may refer to the same disease.
3. Only judge whether the diagnosis result is correct. Information such as the cause of disease, symptoms, and treatment recommendations is not included in the evaluation scope.
4. If the correct diagnosis is included in the predicted diagnosis but some additional complications are mentioned, it is also considered correct.

Output requirements
Only output your judgment result on the model-predicted diagnosis as `Correct` or `Wrong`. Do not output any other content.

Format to follow

```
[Correct|Wrong]
```

Predicted diagnosis:

```
{pred_diagnose}
```

Ground-truth diagnosis:

```
{gt_diagnose}
```

Prompt 41

Prompt for LLM-Judge evaluating treatment-plan accuracy against ground truth on MedRBench test set.

Task description
As a professional medical treatment planning evaluation system, you will receive two treatment plan results for assessment: one is the treatment plan predicted by the model, and the other is the verified correct treatment plan. Your task is to determine whether the model-predicted treatment is accurate.

When evaluating, consider the following factors:

1. If the predicted treatment and ground-truth treatment have exactly the same meaning, then it is correct.
2. If the correct treatment plan is included in the predicted treatment but some additional care is mentioned, it is also considered correct.
3. Considering that even the same disease can sometimes be treated differently, if the model's prediction does not completely match the ground-truth treatment, you can refer to additional information to make a judgment.
4. If the predicted treatment and the ground-truth treatment do not convey the same meaning, and there is no supporting evidence in the additional information to suggest that the predicted treatment is also applicable to the disease, it is considered wrong.

Output requirements
Only output your judgment result on the model-predicted treatment as `Correct` or `Wrong`. Do not output any other content.

Format to follow

```
[Correct|Wrong]
```

Predicted treatment:

```
{pred_treatment}
```

Ground-truth treatment:

```
{gt_treatment}
```

Additional information:

```
{additional_info}
```

References

[1] Xu, R., Wang, Z., Fan, R.-Z., Liu, P.: Benchmarking Benchmark Leakage in Large Language Models. arXiv. arXiv:2404.18824 [cs.CL] (2024). https://doi.org/10.48550/arXiv.2404.18824 . http://arxiv.org/abs/2404.18824 Accessed 2026-08-18

[2] Wu, J., Gu, B., Zhou, R., Xie, K., Snyder, D., Jiang, Y., Carducci, V., Wyss, R., Desai, R.J., Alsentzer, E., Celi, L.A., Rodman, A., Schneeweiss, S., Chen, J.H., Romero-Brufau, S., Lin, K.J., Yang, J.: BRIDGE: benchmarking large language models for understanding real-world clinical practice texts. Nature Biomedical Engineering, 1–16 (2026) https://doi.org/10.1038/s41551-026-01719-2 . Publisher: Nature Publishing Group. Accessed 2026-08-18

[3] Breslow, M.J., Badawi, O.: Severity scoring in the critically ill: part 1–interpretation and accuracy of outcome prediction scoring systems. Chest **141**(1), 245–252 (2012) https://doi.org/10.1378/chest.11-0330

[4] Pellathy, T.P., Pinsky, M.R., Hravnak, M.: Intensive Care Unit Scoring Systems. Critical Care Nurse **41**(4), 54–64 (2021) https://doi.org/10.4037/ccn2021613

[5] Murphy, D.J., Anderson, W., Heavner, S.H., Al-Hakim, T., Cruz-Cano, R., Laudanski, K., Kamaleswaran, R., Badawi, O., Engel, H., Grunwell, J., Herasevich, V., Khanna, A.K., Lamb, K., MacLaren, R., Rincon, T., Sanchez-Pinto, L., Sikora, A.N., Stevens, R.D., Tanner, D., Teeter, W., Wong, A.-K.I., Wynn, J.L., Zhang, X.T., Zimmerman, J.J., Kumar, V., Cobb, J.P., Reuter-Rice, K.E.: Development of a Core Critical Care Data Dictionary With Common Data Elements to Characterize Critical Illness and Injuries Using a Modified Delphi Method. Critical Care Medicine **53**(5), 1045–1054 (2025) https://doi.org/10.1097/CCM.0000000000006595

[6] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**(1), 160035 (2016) https://doi.org/10.1038/sdata.2016.35 . Number: 1 Publisher: Nature Publishing Group. Accessed 2026-06-02

[7] Reese, T., Segall, N., Nesbitt, P., Del Fiol, G., Waller, R., Macpherson, B.C., Tonna, J.E., Wright, M.C.: Patient information organization in the intensive care setting: expert knowledge elicitation with card sorting methods. Journal of the American Medical Informatics Association: JAMIA **25**(8), 1026–1035 (2018) https://doi.org/10.1093/jamia/ocy045